UKERC
UK Energy Research Centre

A report for UKERC by
UKERC Technology & Policy Assessment Function



Energy Efficiency Evaluation:

The evidence for real energy savings from energy efficiency programmes in the household sector

Energy Efficiency Evaluation:

The evidence for real energy savings from energy efficiency programmes in the household sector

A report by the UKERC Technology & Policy Assessment Function

Joanne Wade
Nick Eyre

May 2015

Acknowledgements

This report has been written by Joanne Wade with input from Nick Eyre (Environmental Change Institute, University of Oxford), and supporting research by Victoria Bignet (Environmental Change Institute, University of Oxford). We are very grateful for the guidance and support offered by Rob Gross (Head of UKERC TPA function), Jamie Speirs (UKERC TPA) and Phil Heptonstall (UKERC TPA).

Useful comments and advice were received from our Expert Group and peer reviewers (listed in Appendix A). We thank them for their time and expertise, which has helped greatly. Responsibility for the content of this report, however, remains with the project team.

REF UKERC/RR/TPA/2015/001

www.ukerc.ac.uk
Follow us on Twitter @UKERCHQ

Research Councils UK
Energy
For a Low Carbon Future

Preface

This report was produced by the UK Energy Research Centre's (UKERC) Technology and Policy Assessment (TPA) function.

The TPA was set up to inform decision-making processes and address key controversies in the energy field. It aims to provide authoritative and accessible reports that set very high standards for rigour and transparency. The subject of this report was chosen after extensive consultation with energy sector stakeholders and upon the recommendation of the TPA Advisory Group, which is comprised of independent experts from government, academia and the private sector.

The primary objective of the TPA, reflected in this report, is to provide a thorough review of the current state of knowledge. New research, such as modelling or primary data gathering may be carried out when essential. It also aims to explain its findings in a way that is accessible to non-technical readers and is useful to policymakers.

The TPA uses protocols based upon best practice in evidence-based policy, and UKERC undertook systematic and targeted searches for reports and papers related to this report's key question. Experts and stakeholders were invited to comment and contribute through an expert group. The project scoping note and related materials are available from the UKERC website, together with more details about the TPA and UKERC.

About UKERC

The UK Energy Research Centre (UKERC) carries out world-class, interdisciplinary research into sustainable future energy systems.

It is a focal point of UK energy research and a gateway between the UK and the international energy research communities.

Our whole systems research informs UK policy development and research strategy.

UKERC is funded by The Research Councils UK Energy Programme.

For more information, visit: **www.ukerc.ac.uk**

Executive Summary

This report addresses the question: **What is the evidence that energy efficiency programmes targeted at the household sector have delivered real energy savings?**

Rationale

This report has been written at a time when a more professional evaluation community is developing in Europe, and when household energy efficiency policy in the UK is undergoing a period of substantial change. Policies have been significantly weakened, so that the rate of household energy efficiency improvement has declined substantially and further policy change seems likely. In this context, 'what works' is a timely and important question.

Multiple policies and programmes (actions through which policy is realised) have been employed in the past to encourage improvements in household energy efficiency in the UK and abroad, and many evaluations have been undertaken, but the accuracy of the approaches used has been questioned by some commentators, e.g. Rosenow and Galvin, 2013. The debate between theorists and practitioners about the robustness of existing evaluations, together with the lack of systematic assessment of lessons learned, contributes to uncertainty and controversy over what programmes have achieved and provides an inadequate basis for future programme design. This report aims to improve understanding of what is known, what gaps remain and what current priorities should be for the evaluation community.

Method

The report focuses on energy use within the home, covering programmes tackling any technology relating to in-home energy use and any type of home energy related action. The study reported on here followed procedures established in previous UKERC Technology and Policy Assessment studies, centred on a systematic search of the evidence base of peer-reviewed programme evaluation findings. The restriction to peer-reviewed papers only was intended to help guarantee a minimum level of quality in the evidence reviewed, and was also pragmatic: the significant grey literature on energy efficiency programme evaluations is often not readily accessible, very extensive, and hence impractical to include in a relatively small scale study such as this. However, this does mean this report can only provide a partial view of the current state of knowledge.

The evidence found was assessed using a framework developed during the study, based on theoretical and practical aspects of evaluation good practice.

Evaluation good practice

The purpose of programme outcome evaluation is to estimate as accurately as appropriate the effect of the programme on one or more variables of interest, in this case household energy use. In essence, this requires that the post-programme energy use of a suitably sized sample of households affected by the programme is compared with what this would have been if the programme had not happened (the 'counterfactual'). As the evaluator cannot observe the counterfactual, alternative methods of estimating this have to be found. These methods need to take into account not only the basic effect of the programme on households taking part, but also the effect of any other influences acting on household energy use at the same time, any spillover or rebound effects (resulting reductions or increases in energy use other than those directly targeted), and the extent to which programme participants are 'free-riding' (i.e. would have improved their energy efficiency without the programme). Evaluators also need to take care that the methods used deal adequately with the issue of participants in energy efficiency evaluations being unrepresentative.

The various evaluation methods used to estimate the counterfactual and to compare this with what actually happened each have pros and cons:

- Randomised Control Trials (RCTs) are in theory the most accurate evaluation method for well-defined single interventions on a clearly defined population. However, it may be difficult and expensive to collect all the data required; programme administrators may be unwilling to devote the time and budget required; controlling conditions closely enough to perform a

robust experiment is difficult when complex systems like home energy use are involved, and there may be ethical concerns about providing measures to only some of the households eligible for a programme. Significant improvements in energy efficiency will tend to require large programmes, involving many measures, across diverse populations. Evaluation of these through RCTs is neither feasible nor appropriate.

- Engineering estimates, on the other hand, may be the least accurate evaluation method. However, enhanced estimates that are adjusted to account for known differences between theoretical calculations and practical implementation may offer 'good enough' estimates at relatively low cost, especially for large programmes.
- In between these alternatives are a number of quasi-experimental approaches, each of which has strengths and weaknesses in terms of likely accuracy and difficulty and cost of implementation.

Robust evaluation employs the most appropriate methods for any given situation and, where necessary, uses multiple methods to allow triangulation of results.

Many evaluations are also designed to inform the design of future programmes. This requires understanding not just how participants have behaved, but also why they have behaved in the way they have (including participating at all) and therefore implies the use of some different, and often qualitative methods. Such methods are not relevant to answering our research question about evaluation outcomes, but are an important part of what is considered good evaluation practice.

The evidence base

The literature is widely spread across energy efficiency and evaluation conferences and 20 different journals. It is somewhat dominated by evaluation of programmes undertaken by energy companies, usually as a result of regulatory requirements or incentives. This may indicate a UK/US bias. National, regional and local programmes are all represented in the evidence. In many cases the evaluations were commissioned by the organisation responsible for programme implementation, although many studies did not identify who had commissioned the work. Where papers explained the methods used in sufficient detail, these were generally of a reasonable quality. However, a significant proportion of papers did not provide enough information for their study method quality to be judged. Few papers fully acknowledged the limitations of the evaluation methods used, and very few compared results produced using different evaluation methods.

In terms of methodological good practice, exogenous influences, participant spillover and direct rebound seem generally to be reasonably well accounted for. Free-ridership is less well addressed, and only a small minority of papers clearly addressed the issue of self-selection. Wider effects of programmes (indirect rebound and non-participant spillover) are considered by very few studies.

Findings

The evidence base within the peer-reviewed literature demonstrates a wide range of interesting aspects of the energy saving outcomes of energy efficiency programmes, but the answer to the question posed by this study: *'what is the evidence that energy efficiency programmes targeted at the household sector have delivered real energy savings?'* has to be: in this sub-set of the literature, it is generally affirmative, although partial, varying in quality, and inconclusive regarding the precise magnitude of the energy savings delivered.

Building codes or regulations (energy standards for buildings) form a key part of many Governments' plans for improved home energy efficiency, for example in England and Wales, through the commitment to 'zero carbon new homes' from 2016. Several approaches to estimating the effects of building codes or regulations in a number of different countries are presented in the literature. At the most fundamental level, there is some evidence that building codes do lead to increased energy efficiency, and that they may reduce home heating energy use, but by a smaller amount than ex-ante estimates would suggest. However, the literature offers limited useful quantitative information beyond this.

Building energy labels provide information on the (design or actual) energy use of a building, for example through an Energy Performance Certificate. There is very little evidence in the peer-reviewed literature on the effects of building energy labels, with only two papers identified. These suggest little overall effect of certificates in isolation, but that a significant portion of energy saving potential might be accessed if certificates are provided to people who are already interested in saving energy.

Market transformation activities aim to deliver sustained change in the energy efficiency characteristics of a given market (for example, a type of appliance) through some combination of information, standards and regulation. There is little useful quantitative information regarding the effects of appliance market transformation activities (largely because this type of market transformation activity is difficult to evaluate using the methods commonly employed for energy efficiency programme evaluation). However, results from US Federal standards perhaps give an indication of the level of reduction that might be expected from comprehensive standards programmes covering heating, cooling and electrical appliances: the overall effect could be a reduction in household energy use of just under 10% relative to a 'without appliance standards' baseline.

Programmes of **incentives for investment** form an important part of many Governments' household energy efficiency policies, especially for retrofitting of existing homes. Incentives are generally provided by regulated energy companies, such as the Energy Company Obligation in the UK, or directly by Government, e.g. the former Warm Front programme in England. The peer-reviewed evidence offers no consistent picture of how net direct energy savings in participant households

relate to *ex-ante* engineering estimates for **general investment programmes**: there is a consensus that they are significantly lower, but estimates of the proportion of theoretically possible savings that is achieved range from 44% to 75%. Taking an alternative, macro-level approach, estimating programme impacts using State or national level energy use data and comparing areas with active energy efficiency programmes and those without, provides a different perspective, but no clearer answers. There are a small number of studies in the recent peer-reviewed literature looking at **low-income programmes**, but these are very diverse in their nature and aims, and do not produce an overall picture of likely effects of this type of programme. These are both, however, areas where there is significant additional information in the grey literature that it was not possible to access within the scope of this study.

There is increasing interest in several countries in **innovative finance mechanisms**, such as the UK Green Deal, to reduce reliance on incentives. The only piece of evidence quantifying the outcome of an innovative finance mechanism is from a large and well-established programme in Germany. This supports the idea that the programme is leading to substantial energy savings, but suggests that these may be significantly lower than the programme's own reported average: per household savings may average 25-30% of pre-refurbishment consumption rather than the 54% reported by the programme.

Information and advice programmes are very varied, ranging from basic 'energy saving tips' to detailed in-home advice. The evidence covers a range of approaches, but is skewed towards basic information rather than more in-depth advice. There is very little quantification of effects, and what is presented is not necessarily particularly robust. The methods used here tend to be less robust than for other types of policy, with small sample sizes and reliance on surveys with little reflection on the likely accuracy of responses provided.

Billing feedback is the programme type that has received the most attention in the peer-reviewed literature in recent years. This type of approach has only relatively recently been implemented on a large scale and this, in combination with the availability of smart meter data, has allowed experimental approaches to the study of its effects. Most reports of large scale experimental trials concern programmes implemented in the US, where smart metering is far more prevalent than in Europe, and all but one concern programmes implemented since 2007. The evidence offers a consensus view that feedback programmes result in reductions in household energy use of around one to five per cent.

There is very little robust outcome evaluation of **community-led energy activities** reported in the literature. This may reflect the historical lack of priority given to this type of programme; equally, it could reflect the complexity of objectives in these projects and the preference for the implementers of such actions to focus on process rather than outcome evaluations, aiming to improve initiatives that they already consider to be effective or to increase their reach. As community energy activities evolve to include elements of investment and financial return, evaluation of realised energy use reductions may become more important to the programme implementers.

A review of the literature on the **wider impacts** of household energy efficiency programmes showed that understanding the effects of programmes on non-participants should be areas of concern for evaluators. Recent work suggests that effects that increase energy use (indirect rebound) from commonly implemented household measures and actions may be relatively low, but that effects that reduce energy use (non-participant spillover) from market transformation programmes may be significant. However, more work is needed in both these areas.

In summary, the evidence is relatively clear that:

- Minimum efficiency standards for buildings, appliance market transformation activities, and investment and refurbishment programmes all reduce energy use, although to a lesser extent than ex-ante estimates would suggest. Savings from these types of programme seem to be of the order of 10% of total household energy use for participating households.
- Average effects of feedback programmes are in the range of one to five per cent of participant household energy use although there is a large range around this at the individual household level.

The evidence does not provide clear answers on:

- The likely magnitude of effects such as spillover and free-ridership.
- The outcomes of information and advice programmes other than feedback, community-led programmes, or innovative finance.
- The 'reach' of different types of programme; i.e. the proportion of targeted households that they induce to take action.
- The wider economic impacts of programmes.

The future

There is a need for more consideration of the potential to use Randomised Control Trials, or quasi-experimental alternatives to understand new interventions, both new technologies and behavioural interventions. However, there will remain very good reasons for using theoretically less accurate methods in some circumstances, in particular for cost-effective evaluation of larger and well-understood programmes. More open information and debate about latest understanding of engineering estimates and correction factors, and how these are derived from evaluation evidence, would increase confidence in these evaluation methods.

The limitations of theoretically more accurate methods should also be recognised: Randomised Control Trials

are not appropriate for analysing complex social interventions, they often focus on one fuel only, and hence ignore any programme effects on other fuels used in the home; and the potentially confounding effect of non-participant spillover on methods using comparison groups needs to be acknowledged and taken into account. Increased use of multiple evaluation methods to cross-check evaluation results is one element of this.

There are a number of changes to policy aims that are leading to changes in evaluation design: for example, carbon emissions reduction aims have led to increased investigation of the net, economy-wide effects of energy efficiency programmes. The need to meet economy-wide targets also leads to a requirement to understand the 'reach' of programmes. Most evaluation techniques are designed to measure the energy savings in a specific home and/or the extent to which a programme's objectives have been met. Of more concern in the UK, given recent experience of the Green Deal, is to understand the likely reach of different policy instruments and the aggregate national effect.

One of the most significant changes in programme design in the UK in recent years has been a move away from energy supplier or government funded investment towards mechanisms that support householder investment in energy efficiency. If householders are to be encouraged to invest in energy efficiency on the basis that they will recoup a financial return, it becomes more important to understand how the effectiveness of energy efficiency investment varies between households rather than simply the average effect achieved. The literature reviewed here acknowledges the extent of variation in results, but offers only initial ideas on the determinants of the variance, and therefore is not designed to meet the needs of individual energy efficiency investors.

Another significant change in the UK is increased attention on community energy action. As reported above, there is very little evidence in the peer-reviewed literature on the outcomes of community-based programmes. If the level of ambition for this type of programme is to increase significantly, greater attention to outcome evaluation is urgently needed.

Recommendations for Evaluators

There are three key areas for evaluation research effort:

- Greater understanding of the importance of some of the effects commonly not captured in evaluations (e.g. non-participant spillover);
- Economy-wide impacts of packages of energy efficiency programmes;
- Outcomes of community-led, behaviour change, and innovative finance programmes.

In addition, analysis of the grey literature, and evaluation literature in languages other than English, would contribute greatly to an improved understanding of what we already know.

Evaluation practitioners may want to consider the following issues:

- Greater use of Randomised Control Trials and quasi-experimental alternatives where appropriate, together with more use of multiple evaluation methods to cross-check results;
- Deeper exploration of the variation in effects between different households, making innovative use of the large datasets (e.g. from building energy certification and smart metering) that are now becoming available;
- Greater exposure of evaluation results to discussion in the peer-reviewed literature;
- Presenting evaluation results in such a way that cross-programme comparison is easier (e.g. offering percentage savings figures as well as kWh).

Recommendations for Policymakers

This analysis of evaluations also has a number of implications for policymakers:

- We can say with reasonable confidence that well-established types of energy efficiency programmes can save significant amounts of energy;
- It is now well understood that savings are generally below those that might be calculated by the most basic engineering calculations, and that good design and implementation plans matter because they influence the level of savings achieved;
- Both regulation and incentives programmes can work in different contexts, implying that a range of different energy efficiency policy instruments is needed;
- Some newer types of policy instrument have yet to be as thoroughly evaluated. These include programmes focussed on behavioural change, those led by community groups; and those using new financing mechanisms;
- The key uncertainties in the effectiveness of different policies relate primarily to the scale and reach of policies.

On this basis we can make some reasonably robust recommendations for policy-makers:

- There should be continued support for energy efficiency policies and programmes as these are likely to continue to form cost-effective parts of the delivery of energy policy objectives relating to all aspects of the energy trilemma;
- Well established approaches, such as standards and incentive programmes should form the core of this approach in the short term;
- New approaches may be able to add to the range of policy options, but they need to be piloted and evaluated before there is a commitment to them replacing existing effective approaches;
- Policy makers should take the significant opportunities that exist to learn from experience in other countries and jurisdictions.

Glossary

ACEEE	American Council for an Energy Efficient Economy
Bottom-up	Bottom-up studies are those which estimate energy savings on the basis of changes in each participant household
Building fabric	The basic elements of the building, such as walls, roof, windows etc. Building fabric measures include insulation, double and triple glazing, draught-stripping
CFL	Compact Fluorescent Lamp
Counterfactual	A description of what would have happened to household energy use if the programme had not happened
Demand response	Load shifting away from times of peak demand to enable more efficient use of supply resources
DSM	Demand Side Management: a term used to describe a range of utility-implemented energy efficiency programmes
ECEEE	European Council for an Energy Efficient Economy
EEC	Energy Efficiency Commitment: a UK energy supplier obligation to deliver energy efficiency investment
ECO	The Energy Companies Obligation: the latest energy efficiency obligation placed on UK energy companies
EnergyStar®	A US federal energy efficiency labelling programme
Energy service	A service requiring the use of energy, for example a warm home. Householders demand energy services and this leads to a demand for electricity, gas and other fuels
Euclidean distance	The shortest distance between two points. In this context, it is the difference between two values of a variable affecting energy use, and is used in matching households for quasi-experimental estimates of programme effects.
EUL	Effective Useful Life: a factor used in calculations of programme outcomes to reflect the extent to which measures remain in use / fully effective as time progresses
Ex-ante	Before: refers to estimates of programme effects before the programme has been implemented
Ex-post	After: refers to estimates of programme effects based on data gathered during and after programme implementation
Exogenous	Outside: describes factors that influence household energy use other than the programme
Free-ridership	A measure of the extent to which households participating in a programme would have taken the energy efficiency actions promoted by the programme even without the programme's incentives
Fuel poverty	Inability of a household to afford to maintain their home at a temperature that promotes health and to afford other basic energy services
GJ	Gigajoules
GWh	Gigawatt hours
Heat replacement effect	The increase in home heating required when more efficient lighting or appliances are installed (greater efficiency means less waste heat)
IEA	International Energy Agency

IEPEC	International Energy Program Evaluation Conference
kWh	Kilowatt hour
Measures	Technologies that increase energy efficiency (for example, loft insulation or an efficient refrigerator)
Net savings	The proportion of observed energy savings that can be attributed to a programme. Precise definitions vary between programmes, but the term is often used to denote an adjustment to account for free-ridership.
Non-participant	Refers to a household identified as not taking part in a programme. Non-participant households are used to provide a comparison group to participants for use in quasi-experimental evaluations
Participant	Refers to a household identified as taking part in a programme
Persistence	The extent to which the energy saving outcome of a programme is maintained over time
Policy	A set of government aims and objectives linked to a specific issue
Portfolio	A group of projects implemented by a single organisation
Prebound	The extent to which household energy use prior to programme implementation differs from a calculated value
Programme	Any set of practical actions through which a policy is realised
Propensity score matching	Matching participant and comparison group households based on the likelihood of their participation in the programme
Rebound	Increased energy use resulting from the reduced cost of providing a given level of an energy service following investment in increased energy efficiency. This can offset some of the expected energy savings from energy efficiency measures.
Rigour	The degree of accuracy and precision of a method / result
Self-selection bias	A potential bias in evaluation results introduced by characteristics (observable and unobservable) of participant households that influence their likelihood of participating in a programme and also their reaction to both the programme and to exogenous variables that affect energy use
Standard Assessment Procedure	A method of calculating the energy efficiency of a home. This method underlies energy ratings and energy certificates for homes in the UK.
Spillover	Energy use reductions that happen as a result of a programme but that are not directly supported by the programme
Top-down	Top-down methods involve estimation from macro-level data such as changes in total household energy use in a given geographical area
TWh	Terawatt hours
UKERC	The UK Energy Research Centre
Weather correction	An adjustment to calculated or observed differences in energy use at different times to account for any change in temperature between the two time periods
Weather normalisation	An adjustment to calculated or observed differences in energy use to account for differences between temperatures during the time periods being compared and the average annual temperatures for the location in question

Contents

Introduction

1

1. Introduction

The UK Energy Research Centre (UKERC) Technology and Policy Assessment (TPA) function was set up to address key controversies in the energy field through comprehensive assessments of the current state of knowledge. It aims to provide authoritative reports that set high standards for rigour and transparency, while explaining results in a way that is both accessible to non-technical readers and useful to policymakers. This latest report addresses the following question:

> *What is the evidence that energy efficiency programmes targeted at the household sector have delivered real energy savings?*

1.1. Rationale

Improvements in household energy efficiency are a necessary element of the transition to a sustainable energy system, to meet climate change and social policy objectives. Delivering such improvements is a long-standing element of energy policies in many countries and in the UK is the focus of important new policy initiatives (notably the Green Deal[1]). Multiple policies and programmes[2] have been employed in the past to encourage such improvements, and many evaluations have been undertaken[3]. However, the rigour of these evaluations has been questioned, and the lack of systematic assessment of the lessons learned highlighted (Frondel and Schmidt, 2005).

There is an extensive grey literature on energy efficiency programme evaluation (notably of utility programmes in the US[4]). However, there is a perception that many of the evaluations reported rely on an engineering approach to savings estimation, and the accuracy of this type of approach has been questioned (Gillingham et al., 2006). Evaluation literature proposes that good quality evaluation requires either careful econometric analysis of aggregate data (Horowitz, 2007) or experimental or quasi-experimental studies, based upon accurate before and after measurement of energy consumption and controlling for factors such as selection bias and free-ridership (Hartman, 1988). However, energy efficiency programmes are implemented within a complex socio-technical system, and evaluations are conducted on real programmes within budget and data constraints. These conditions place limits on the applicability of theoretically optimal evaluation techniques.

The debate between theorists and practitioners about the robustness of existing evaluations, together with the lack of systematic assessment of lessons learned, contributes to uncertainty and controversy over what previous energy efficiency programmes have achieved and provides an inadequate basis for future policy design. This TPA assessment of the evidence offered by existing evaluations of energy-efficiency programmes is intended to improve understanding of what we know about 'what works' in household energy efficiency policy, where there are gaps in this knowledge and what can be done to improve our understanding.

1.2. Context

This study of household energy efficiency programme evaluation is being conducted at a time when 'Europe is at a critical juncture in developing a professional [energy programme] evaluation community' (Vine and Thomas, 2012). This suggests that the quality of previous European programme evaluations, and their reporting, may be variable. Vine and Thomas point to the US as the location with the most experience of programme evaluation to date but even there, as Davis et al (2013) note: 'although many studies have evaluated the effectiveness of such interventions, their designs and reporting protocols vary so much that it is hard to aggregate their results'. This presents a challenge to this study and may limit the extent to which quantitative estimates of programme outcomes can be determined.

It is also being conducted at a time when household energy efficiency policy in the UK is undergoing a period of substantial change. Government funded fuel poverty programmes in England have been ended and energy supplier programmes have been significantly weakened, so that the rate of household energy efficiency improvement has declined substantially (Rosenow and Eyre, 2013). Critical assessment by independent analysts and cross-party parliamentary committees seems likely to lead to policy change, whatever the outcome of the 2015 General Election. For the analysis of the evidence base, this study has grouped programmes into a number of key areas. Box 1.1 describes briefly the main elements of current UK energy efficiency action within each of these areas.

1 See Box 1.1 for more on the Green Deal.

2 For definitions of 'policy' and 'programme', see section 1.3.

3 See for example the MURE database (http://www.measures-odyssee-mure.eu/) for headline results from a wide range of programmes implemented in European countries.

4 For example, see http://www.calmac.org/

Box 1.1 UK energy efficiency actions

Minimum efficiency standards for buildings: Minimum energy efficiency standards have been included in the UK Building Regulations since 1965. In 2006 the level of ambition in the regulations was increased significantly with the announcement that all new-build housing would be required to be net zero carbon[5] from 2016; an aim that is being implemented through successive tightening of the energy efficiency standards within the regulations (CLG, 2007). This is in advance of European requirements for 'nearly zero carbon' buildings by 2020 (EP, 2010).

Energy labelling for buildings: Energy Performance Certificates for homes were introduced in 2007 in compliance with European legislation on the energy labelling of buildings (EP, 2002, EP, 2010). These are based on the UK Standard Assessment Procedure for calculating the energy efficiency of a home, and include information on running costs, current and potential energy efficiency levels, and recommended energy efficiency measures.

Appliance market transformation activities: EU energy labelling and minimum efficiency standards for domestic appliances have been implemented gradually since 1995. Current activity is within the framework of the EU Ecodesign Directive[6] (EP, 2009) and centres on information and minimum efficiency standards, with no significant incentive programmes operating.

Large scale investment and refurbishment programmes: Historically, the UK government has supported investment in home energy efficiency through a combination of taxpayer funded schemes for low-income households and regulatory requirements on energy companies. Obligations to invest in energy efficiency on behalf of domestic customers have been placed on electricity and gas suppliers since Energy Efficiency Standards of Performance were introduced in 1994[7]. Schemes to deliver the obligations have included direct installation of insulation and heating measures, discounts on DIY measures and appliances and the provision of CFLs. The latest set of obligations (the Energy Companies Obligation, ECO), were introduced in January 2013 and are intended to work alongside the Green Deal. As with previous obligations, larger energy companies are required to deliver energy efficiency measures to householders. ECO includes a Carbon Emissions Reduction obligation (subsidising measures that cannot be funded entirely through Green Deal finance – see below - because they are not sufficiently attractive to consumers); a Carbon Saving Communities obligation (delivering insulation and connection to district heating networks in low-income areas); and a Home Heating Cost Reduction obligation (providing insulation and heating measures to low income and vulnerable households). The low-income elements of ECO are now the only national funding supporting investment in the homes of low-income households in England, as the taxpayer funded 'Warm Front[8]' scheme ceased in January 2013. However, national government funding for energy efficiency measures for low-income households remains in place in Wales, Scotland and Northern Ireland.

Innovative finance: New options for funding home energy efficiency investments are being developed in many countries. The UK version, the Green Deal[9], was officially launched in early 2013, and is often described as the main UK Government policy to support energy efficiency investment in the household sector[10]. It centres on the provision of household specific advice, through a Green Deal Advice Report, and the availability of finance for energy efficiency investments that is repaid through a charge on the electricity bill. The finance is at market rates of interest, is provided only if the investments are sufficiently cost-effective that energy bill savings are equal to or greater than finance repayments, and responsibility for repayment remains with the property on change of ownership or tenancy.

Information and advice: Until relatively recently, the UK government supported the provision of telephone and face to face energy advice to householders through a network of Energy Efficiency Advice Centres. As part of the implementation of the Green Deal, this advice provision has now been centralised into a telephone and online service[11], although in some local areas, advice centres continue to be supported, for example by local government.

Smart metering: the UK is introducing smart metering in line with European regulatory requirements (EP, 2012). The rationale for introducing metering in the UK is not primarily linked to demand response (load shifting) but rather to the opportunity to induce energy savings through feedback mechanisms (Darby et al., 2011).

Community-led energy action: successive national Governments have recognised the importance of energy action at the local and community level in the transition to a low carbon economy (Wade et al., 2013). Various government funding competitions have supported pilot actions, led by local authorities and community organisations. As a result of this and other, local, drivers there is now a patchwork of community energy activity across the country. Much of this is focused on community ownership of renewable energy resources, but some includes action on household energy use. The Government recently formalised its framework for this type of action in a Community Energy Strategy (DECC, 2014).

1.3. Key definitions

There are inconsistencies and overlaps in the terminology used in the literature about energy efficiency programme evaluation. Here a number of key terms are defined to clarify the use of these terms in this report.

Policy and programme terminology

The terms 'policy' and 'programme' are used differently, and sometimes interchangeably, by different authors. Their meaning in this report is defined here, together with the meaning of a number of related terms.

Policy

Policy is used in this report to refer to a set of government aims and objectives linked to a specific issue. For example, it is government policy to reduce UK carbon emissions in line with the carbon budgets set out subsequent to the Climate Change Act 2008[12].

Programme

Programme is used to refer to any set of practical actions through which a policy is realised. This is a broader definition of the word than is generally employed: the usual large scale refurbishment or incentive programmes, such as those implemented by utilities, are included in this term; but also actions such as the introduction of minimum efficiency standards for buildings or appliances.

The distinction between policy and programme is illustrated further in Table 1.1.

Table 1.1: policy and programme examples

Policy	Programme
Reduce carbon emissions from new buildings to contribute to meeting Climate Change Act emissions reduction targets	Building regulations
Support the development of a market for more energy efficient homes to contribute to carbon emissions reductions	Energy Performance Certificates
Reduce energy use by domestic electrical appliances	The appliance market transformation programme
Address market failures by requiring energy efficiency investments by energy suppliers	The Energy Companies Obligation
Reduce the number of households in fuel poverty	The Energy Companies Obligation
Encourage increased householder investment in energy efficiency measures to contribute to carbon emissions reductions	The Green Deal
Encourage increased householder action on energy efficiency to contribute to carbon emissions reductions and fuel poverty alleviation	The Energy Saving Advice Service
Improve householder information on energy use to encourage better household energy management	Smart meter roll out (and billing feedback)
Increase engagement in energy efficiency action through the Community Energy Strategy	Peer to peer mentoring for new entrants from experienced community energy groups

5 The definition of 'net zero carbon' is yet to be finalised, but it will include requirements for cost-effective energy efficiency and on-site renewable energy measures, together with investment in off-site measures where on-site measures are insufficient to reduce net carbon emissions from the building to zero.

6 The Ecodesign Directive sets a framework that allows the use of regulations at the European level that require mandatory ecodesign elements for some energy-related products (see http://ec.europa.eu/enterprise/policies/sustainable-business/ecodesign/index_en.htm for more information).

7 https://www.ofgem.gov.uk/environmental-programmes/energy-companies-obligation-eco

8 Warm Front was a national scheme offering heating and insulation measures to low-income householder in receipt of certain income-related benefits.

9 A more detailed description of the Green Deal can be found here: http://www.energysavingtrust.org.uk/Take-action/Find-a-grant/Green-Deal-and-ECO

10 The Green Deal also covers non-domestic buildings.

11 http://www.energysavingtrust.org.uk/Organisations/Government-and-local-programmes/Programmes-we-deliver/Energy-Saving-Advice-Service

12 http://www.legislation.gov.uk/ukpga/2008/27/contents

Project

Programmes may comprise more than one project (for example, an energy supplier may implement a series of projects under the Energy Companies Obligation). Evaluations of individual projects are included in this study where they offer insight into the effectiveness of the programme they are part of.

Portfolio

A series of projects implemented by one organisation (for example, an energy supplier's activity under ECO) may be referred to as a portfolio. This term is most often used when referring to utility activities, but may equally be applied to activities by other organisations, including local or national government.

Mechanism

One project or programme may employ more than one mechanism to encourage energy efficiency actions. For example, the appliance market transformation programme includes energy labelling, minimum efficiency standards and a number of additional education and information initiatives for retailers and consumers. In some cases, the effects of each mechanism are evaluated separately; in others the evaluation focuses on the overall effect of the project or programme without attempting to allocate this between the various mechanisms employed.

Measures and energy-use actions

A project or programme may focus on a single energy efficiency measure (for example, energy efficient boilers) or a range of measures (for example, all cost-effective heating and insulation-related investments). Projects and programmes may also focus on changing energy-use actions in the home (for example, encouraging clothes washing at a lower temperature). The generic term 'measures' is sometimes used to denote both energy efficiency technologies and efficient energy-use actions, but in this report the two are kept separate, although it should be noted that technologies and behaviour cannot always be separated, since each influences the other.

Evaluation

The Oxford English Dictionary defines evaluation as: The making of a judgement about the amount, number, or value of something; assessment.

At a high level, evaluations split into two groups, depending on their aims:

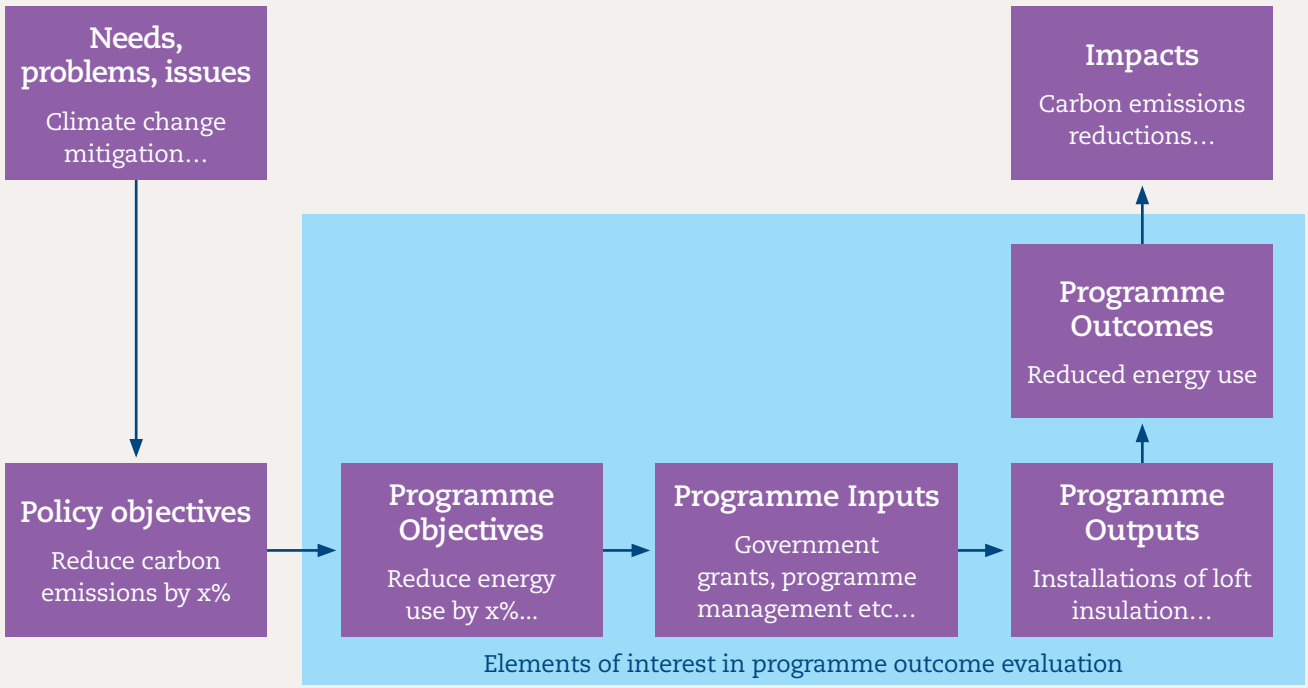
- Impact or outcome evaluations (estimating the programme effects on e.g. energy use, greenhouse gas emissions, customer retention, the overall market for the fuel affected).
- Process evaluations (determining how the process of implementing the programme could be improved).

This study is interested in the extent to which programmes result in energy savings; hence it is concerned primarily with impact or outcome evaluations, whilst recognising that evaluations may have other legitimate goals.

Impact or outcome evaluation of policies and programmes can have a number of objectives (Vreuls, 2005): understanding the extent to which impacts address the needs that the policy was aiming to meet (the utility of the policy); understanding the extent to which the programme's outcomes meet its objectives (the effectiveness of the programme); understanding the level of outputs generated by a given level of inputs (the programme's efficiency); and understanding the extent to which the objectives of the programme are consistent with the identified needs (the programme's relevance).

This study is interested in the extent to which energy efficiency programmes deliver their objective of reduced energy use; hence it is concerned with evaluations of programme effectiveness, looking at programme outputs and outcomes in relation to their inputs and objectives. Figure 1.1 (adapted from Vreuls, 2005) illustrates how the evaluations considered here relate to the policy process as a whole.

Figure 1.1: programme outcome evaluation



In this report, evaluation is taken to include all methods that are used to make an ex-post[13] judgement on the amount of energy saved by a programme. This includes, at its simplest level, calculations based on previous engineering estimates of energy savings from a given measure combined with knowledge of the actual number of measures installed during a programme. It excludes ex-ante estimates of programme effects.

Household energy use

This study is interested in the total use of energy within homes. Hence it treats each unit of energy consumption equally, regardless of the fuel supplying the energy. We consider this appropriate for a study of energy savings, whilst recognising that it would not be sufficient if the focus was on carbon emissions reduction, for example.

Note that the total use of energy within a home is not necessarily the same as energy use measured as the energy flow across the property boundary, since some households will meet a portion of their energy demand from on-site micro-generation technologies. Some studies will define energy use as the actual use of energy within the home, others as the flow across the boundary through the meter. Where programmes involve the promotion of micro-generation technologies, care has been taken to use only data for use of energy in the home in this study.

Programme effects terminology

To accurately estimate the effect of an energy efficiency programme on energy use in homes, a number of effects must be taken into account within the calculation methodology. These are defined here and their potential effect on estimates summarised in Figure 1.2.

Exogenous influences

These are factors other than the programme that may affect energy use in the home (for example, energy price changes). When these act at the same time as the programme, the effect of the programme has to be separated from the effects of these exogenous influences.

Rebound

If a programme results in a household using less energy to gain the same energy service (for example, to heat their home to the same temperature for the same amount of time), a number of rebound effects may occur, and each of these will lead to lower energy savings than might be expected. Total rebound is a combination of direct and indirect effects.

Direct rebound occurs when the householder uses more of an energy service targeted by a programme, as a result of that programme. For example, following the installation of insulation, a householder may heat their home to a higher temperature, or for a longer period of time, because the cost of this has reduced or because the home warms up more easily.

13 'Ex post' is after the programme has been implemented, whilst 'ex ante' is before programme implementation.

Indirect rebound is the combination of a number of effects. Firstly, a householder saving money because a programme has reduced the amount of energy needed to supply a service, may choose to spend some of this money on other home energy services, thus increasing energy use in the home for these services. Secondly, they may choose to spend some of the money on goods and services that result in increased energy use outside the home. Thirdly, the manufacture of any energy efficiency measures supported by the programme will result in energy use[14]. And finally, there are a number of longer term economic effects of the increased energy efficiency, such as impacts on the price of energy and thence on the relative prices of energy intensive goods and services, which will in turn affect energy use.

Spillover

A programme may result in energy use reductions other than those it directly supports; this effect is known as spillover.

Participant spillover occurs when households participating in a programme take additional energy use reduction actions. For example, a householder receiving subsidised cavity wall insulation may choose to also install loft insulation without any support from the programme. Equally, a householder may purchase a number of subsidised Compact Fluorescent Lamps (CFLs) and then decide to buy more CFLs without the subsidy.

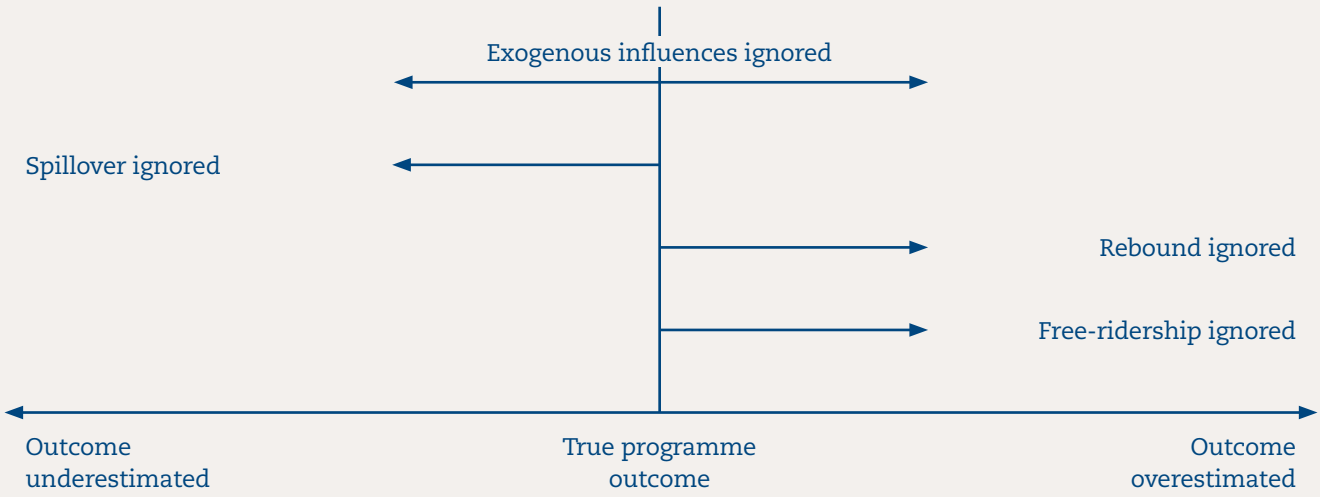
The additional action may affect the same energy service as affected by the programme, as in the two examples given, or it may affect another energy service within the home: for example, a programme may promote energy efficient refrigerators and a participant householder may decide to purchase a new, more efficient fridge and also a new, more efficient washing machine.

Non-participant spillover occurs when households take energy saving actions as a result of the programme but without officially participating in it. For example, a householder may see a programme offer of a rebate on an energy efficiency measure and decide on the basis of this to invest in the measure but then not actually claim the rebate. Larger scale programmes may also have wider non-participant spillover effects via changes in the market for energy efficiency options: for example a programme may increase sales of a measure to such an extent that the price is reduced; this will tend to encourage non-participants to purchase the measure.

Free-riders

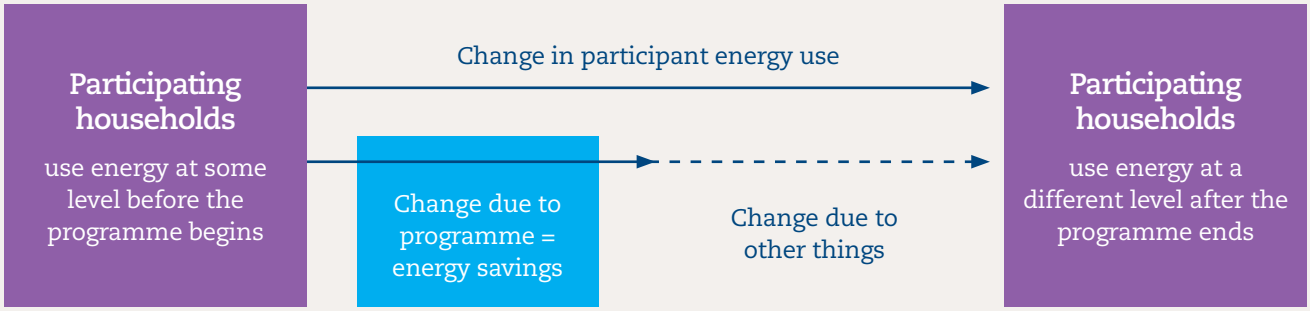
A proportion of the people who take action, seemingly as a result of a programme, would have taken this action even in the absence of the programme. For example, someone planning to purchase a new, energy efficient fridge may claim a rebate from a programme even if they would have purchased the same fridge without the incentive of a rebate. These people are known as free-riders.

Figure 1.2: impacts on savings estimates



14 This 'embodied energy' is often treated separately from rebound, but it is included here because the estimates of rebound levels discussed later in the report include it.

Figure 1.3: energy savings by participant households



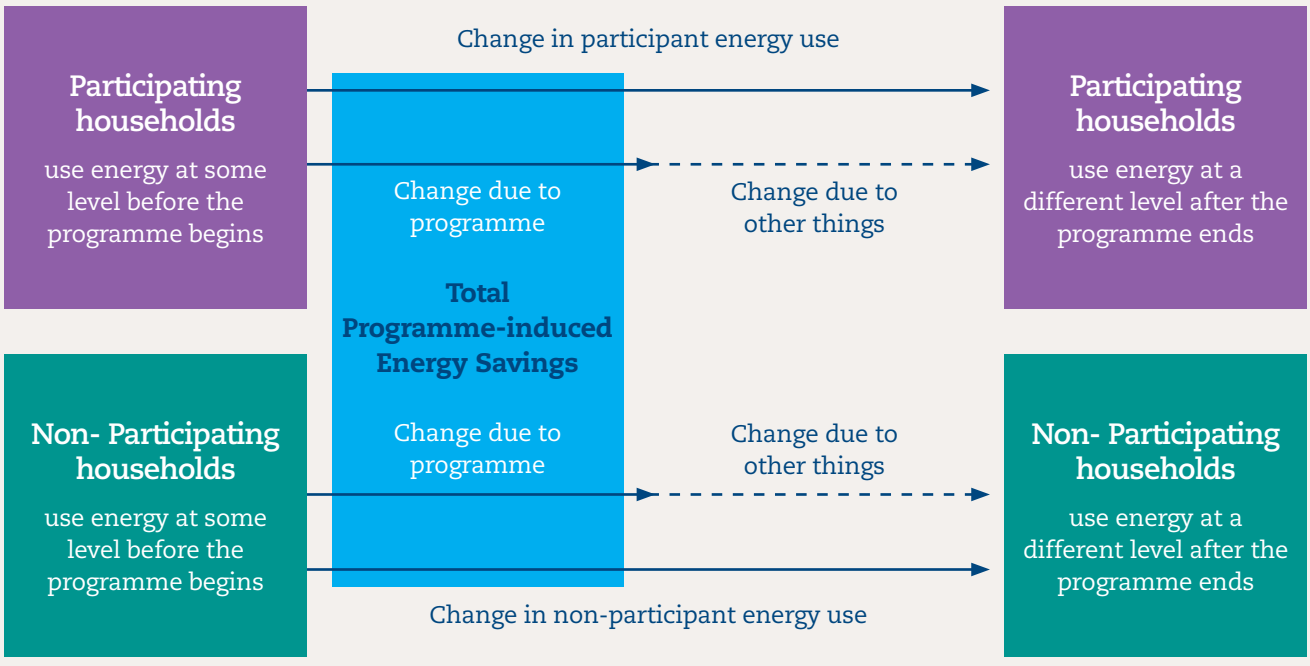
Energy savings

'Energy savings' refers to the reduction in energy use in homes resulting from the programme being evaluated. The primary focus in this study is on energy savings by households participating in the programme. It can be described as the observed change in energy use in the homes of participant households, less any change that is not caused by the programme. Figure 1.3 illustrates this.

It includes the initial effects of the programme on the energy end-use it targets, any direct rebound and participant spillover. It also includes the elements of indirect rebound that affect energy use in the home. Changes in energy use in the targeted households that happen at the same time but are not the result of the programme are excluded, as are changes that may appear to be due to the programme but can actually be assigned to free-ridership. The quantity of concern in this report is therefore consistent with one commonly used definition of 'net savings'[15].

Total energy savings as a result of a programme also include any non-participant spillover[16]. Figure 1.4 illustrates total programme-induced energy savings.

Figure 1.4: total programme-induced energy savings



15 Definitions of net savings vary between different programme evaluations, which can present difficulties for reviews such as this one since it renders comparisons across studies difficult if not impossible. This issue is discussed further in the findings chapter of this report.

16 Note that an understanding of likely non-participant spillover is also important for some methods of estimating changes in participant energy use, because they rely on a comparison with a group of non-participants. This issue is discussed further in Chapter 2.

In the UK, the ongoing political arguments about the consumer benefits of energy efficiency programmes suggest that some doubt about their cost-effectiveness remains, and it can be argued therefore that evaluation of individual programme outcomes in participant households is still important: hence this is the primary focus of this study.

However, an increasing number of programmes aim to transform the market for energy efficient products and/or change social norms related to energy using actions. These programmes can affect decisions made by many people other than those who can be easily identified as programme participants (for example, by altering the choice and price of appliances on the market, or by raising general awareness of energy efficiency as an issue). Hence, understanding the outcomes for programme participants is clearly not sufficient in itself. Therefore, in discussing the current state of knowledge this study does also consider briefly the extent to which programme evaluations provide information about energy saving outcomes beyond those achieved amongst programme participants (including non-participant spillover and the full effect of indirect rebound), and the extent to which alternative approaches (eg market effects studies and econometric modelling of macro level effects of portfolios or policies) are beginning to address this issue.

1.4. Scope of the literature review

The field of 'energy efficiency programmes targeted at the household sector' is a broad one and therefore it is important to define the boundaries of this review clearly so that the reader can understand what has been feasible within the limits of this one study.

Type of literature

The evidence reviewed for this study was restricted to peer-reviewed papers only[17]. There is a very significant body of evaluation work outside of this[18] but this has been excluded for three reasons: first, it is often not readily accessible through key databases or conference papers[19] - a systematic review would be extremely time consuming to develop and implement and is therefore not possible within the scope of this relatively short review study; second, the process of peer review should, in theory at least, help to guarantee a minimum level of quality in the materials reviewed; and third, it was hoped that the peer reviewed literature contains discussion of evaluations of most of the major programme types. The extent to which this restriction to peer-reviewed papers has biased or limited the results of the study is discussed in Section 4.10.

Sectors

The purpose of the study was to look at programmes for household energy efficiency. Therefore, the primary focus of the review was on literature specifically about the household sector. However, where literature covering a broader range of sectors is sufficiently relevant to prove useful for the review, it has been included.

Geography

The literature survey was initially global (but restricted to publications in English). However, one element of the assessment of relevance was whether or not the results of the evaluation might be applicable to the UK situation. Hence the literature forming the evidence for this report is largely concentrated in Europe and North America, with a limited number of papers reporting evaluations elsewhere in the world[20].

Programmes

All types of energy efficiency programme that support increased uptake of existing energy efficient technologies, or changes in how these are used, are included in the review[21]. For analysis purposes, programmes are grouped as in Box 1.1 (minimum efficiency standards for buildings; energy labelling of buildings; appliance market transformation activities; large scale investment and refurbishment programmes; innovative finance; information and advice; smart metering and billing feedback; and community-led energy action).

Route to reducing energy use

Programmes may affect energy use through changes in available technologies, changes in technology purchase decisions and changes in the way new and existing technologies are used in the home. This review encompasses programmes targeting all technologies relating to household energy end-use in the home[22] (residential buildings and their fabric insulation; space conditioning systems and controls; water heating systems and controls; lighting; and major household electrical

17 No distinction is made between the level of peer review for journal papers and that for conference papers as the conferences included here do require peer review of full papers.

18 For example, the Calmac website (www.calmac.org) provides information on 208 impact evaluation reports for residential sector energy efficiency programmes in California alone whilst the US Energy Information Administration's evaluation inventory contains 329 data sources for energy efficiency program evaluation.

19 Evaluation reports in Europe are not as easily accessible as those in the US: the ODYSSEE-MURE database for example (http://www.odyssee-mure.eu/) reports on programme impact estimates in National Energy Efficiency Action Plans, but these are often ex-ante projections, and there are no links given to ex-post evaluation reports. An additional issue for European programmes is that original evaluation reports will generally not be written in English and hence would only be accessible to a multi-lingual review team.

20 A full analysis of the geography of the evidence base is contained in chapter 3.

21 Programmes of support for development of new technologies are not included as these present very different evaluation challenges. Economy-wide policies such as energy taxation and emissions trading are also excluded, for similar reasons.

22 Household energy use for transport is excluded from this study.

appliances) and all types of energy-related actions (longer-term one-off actions such as the purchase of insulation; occasional actions such as changes to heating system controls; and every-day or habitual actions such as the choice of programme on the washing machine).

Energy use in the home is affected by technologies and behaviours associated with water use: policies and programmes to reduce household water use will therefore have a direct impact on household energy use also, relating to the proportion of water that is heated before use. However, these policies are outside the scope of this study since their primary aim is not to increase energy efficiency.

Policies and programmes that aim to change broader social practices (for example food preparation and consumption) will have an impact on home energy use. However, as for water demand management policies, increased energy efficiency is not their aim and hence they are outside the scope of this study.

1.5. Study Method

This study has followed procedures established in previous TPA assessments. It began with the definition of the project in a scoping note, which was published on the UKERC website[23]. An expert group was convened to advise the project team, representing a variety of opinions and perspectives. The members of this expert group are listed in Appendix A. Two meetings of the group were held during the course of the study: one to discuss the project scope and the other to discuss emerging findings. Expert group members were also given the chance to provide feedback on the draft final report.

A systematic search of the evidence base of peer-reviewed programme evaluation findings was carried out. This covered key databases and also proceedings from key peer-reviewed conferences. The databases, search terms used, and conference proceedings included are detailed in Appendix B. Note that the conference proceedings are not searchable in the same way as the main databases and hence the complete proceedings were reviewed, and papers that appeared relevant based on their abstract were added to the initial list of papers for review.

Once the database search returns had been cleaned (duplicate and clearly irrelevant returns removed), references were stored in an Endnote library and categorised, based on reading the abstract, as: priority (focus is clearly on the quantitative or qualitative results of household energy efficiency programme evaluation); context (focus includes elements of household energy efficiency programme evaluation, but is not on presentation of results of evaluations); methodology (focus is on evaluation methodologies); and exclude (focus does not appear to be directly relevant to this study).

The 'methodology' papers were drawn upon to help develop a description of evaluation good practice: both theoretical and pragmatic. This description was used to develop a framework for the assessment of the literature reporting evaluations of programmes. This description and framework form the next chapter of this report.

The 'priority' literature was then categorised (according to programme type, where and when implemented, who implemented the programme, who commissioned the evaluation, and what evaluation methods were used) and reviewed using the assessment framework developed during this project and described in Chapter 2, below. 'Context' literature was drawn upon as appropriate during the discussion of the evidence contained in the 'priority' literature. The findings of this review form the third chapter of this report. Note that the limited time available for the project did not allow for questions to be asked of paper authors, therefore the analysis presented here is based on the written material in the papers, with no further clarification from their authors.

The draft report was peer reviewed. The peer reviewers are listed in Appendix A. Their comments, together with those from the Expert Group, were considered and this final report produced.

1.6. Report structure

The next chapter of this report describes elements of evaluation good practice from both theoretical and pragmatic perspectives, and explains the assessment framework used in the review of the programme evaluation literature.

Chapter 3 provides an overview of the literature reviewed in this report, addressing the types of programme covered and the quality of the evidence based, as assessed within the framework developed in Chapter 2.

Chapter 4 presents the study findings on the effect of household energy efficiency programmes on household energy use. It is structured according to the programme types defined earlier in this introduction. The results for each programme type are followed by a discussion of key findings from the evidence base as a whole.

Chapter 5 concludes the report with a series of recommendations concerning evaluation theory, evaluation practice and priorities, and the need for analysis of the evidence base outside the peer-reviewed literature.

23 http://www.ukerc.ac.uk/programmes/technology-and-policy-assessment/energy-efficiency-evaluation.html

2. Evaluation good practice

This chapter examines how 'good practice' can be defined for energy efficiency programme evaluation. It begins with consideration of the theory supporting programme evaluation. Following this, the main methods used to evaluate programmes are described, and the extent to which each method can, in theory, accurately estimate programme effects is considered. A summary comparison of the methods is presented, highlighting key benefits and drawbacks of each method and noting when each may be used. The chapter then moves on to consider evaluation in practice, examining some of the challenges facing evaluators and exploring how these influence the choice of evaluation method. The chapter concludes with an explanation of the framework used in this project to assess the literature, based on the review of evaluation theory, methods and practicalities described in the preceding sections.

2.1. The theory behind good evaluation

The purpose of a programme outcome evaluation is to estimate as accurately as appropriate (given data, time and budget constraints) the effect of the programme on one or more variables of interest, in this case household energy end-use. In essence, this requires that the post-programme energy use of a suitably sized sample of households affected by the programme is compared with what this would have been if the programme had not happened (the 'counterfactual').

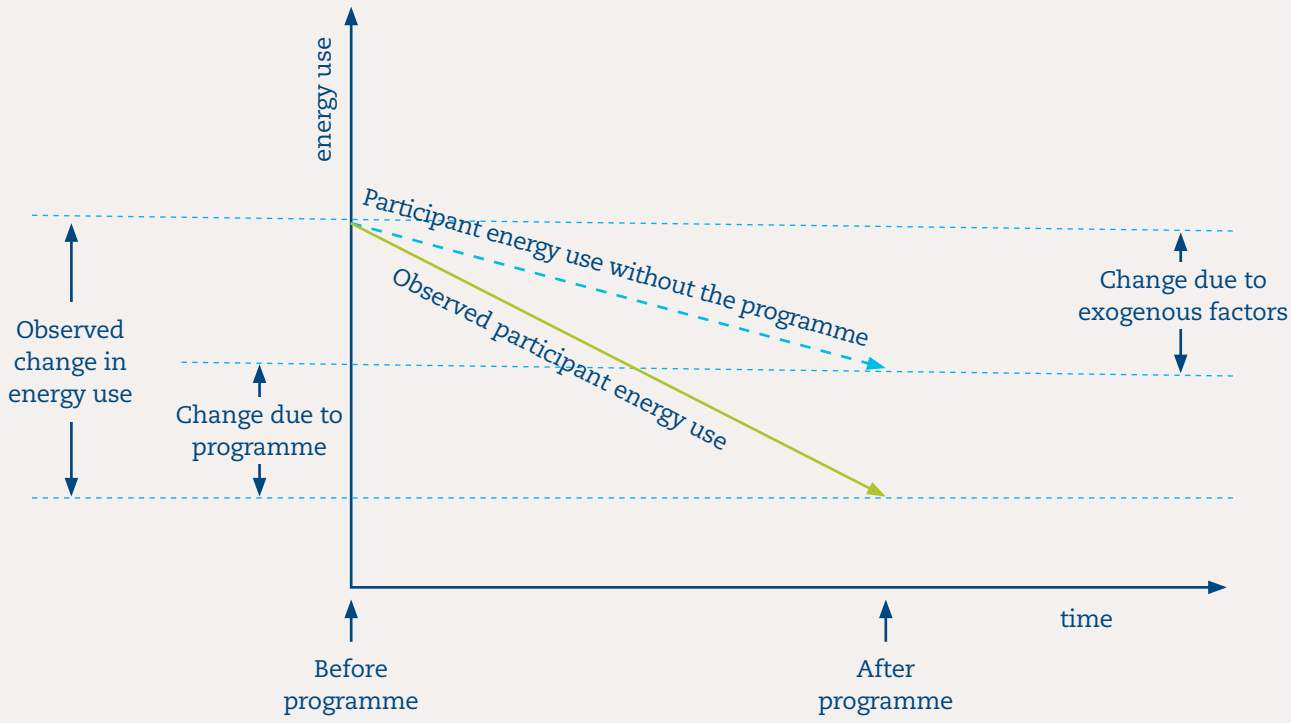
The most obvious issue with this is that the evaluator cannot observe how much energy would have been used by the affected households if the programme had not happened. Hence an alternative way to estimate this counterfactual has to be found (Frondel and Schmidt, 2005).

Defining a counterfactual

Assuming to begin with that the households of concern to the evaluation are only those that have explicitly participated in the programme and that the evaluation is interested in the effect of the programme on total home energy use, one simple estimate of the counterfactual is the energy use of participant households before the programme was implemented. Using this method, if the total home energy use before and after the programme is measured, the evaluation will capture the effect of the programme on the particular technology or energy using action that has been targeted. It will also capture any participant spillover, direct rebound effects and the proportion of indirect rebound affecting energy services within the home[24].

However, this is only accurate if there are no other factors acting to affect participant energy use between the 'before' and 'after' energy use measurements. This is very unlikely to be the case. Figure 2.1 illustrates the effect of exogenous variables.

Figure 2.1: exogenous variables and before-after comparisons



24 The method will result in these effects being accounted for within the overall estimate of the effects of the programme; it will not separately identify or quantify them.

Some of the exogenous influences may be relatively easy to adjust for (e.g. changes in the weather that affect heating energy use) but others will be more difficult (e.g. changes in income or prices, autonomous technological change, or responses to other energy saving initiatives running at the same time). Hence, if it is likely that other factors may be causing changes in energy use, comparing the energy use of participant households with that of non-participants may provide a more accurate estimate of programme effects[25].

The evaluation problem then becomes the selection of an appropriate group of non-participants to provide this comparison group. For an accurate estimate, the comparison group needs to be identical to the participant group in every way other than the fact of participation in the programme. Evaluators can match participant and non-participant groups as closely as possible based on observable characteristics such as socio-economic data, taking into account how these are known to interact with other determinants of energy end-use. However, two issues remain, which together are known as the self-selection bias problem.

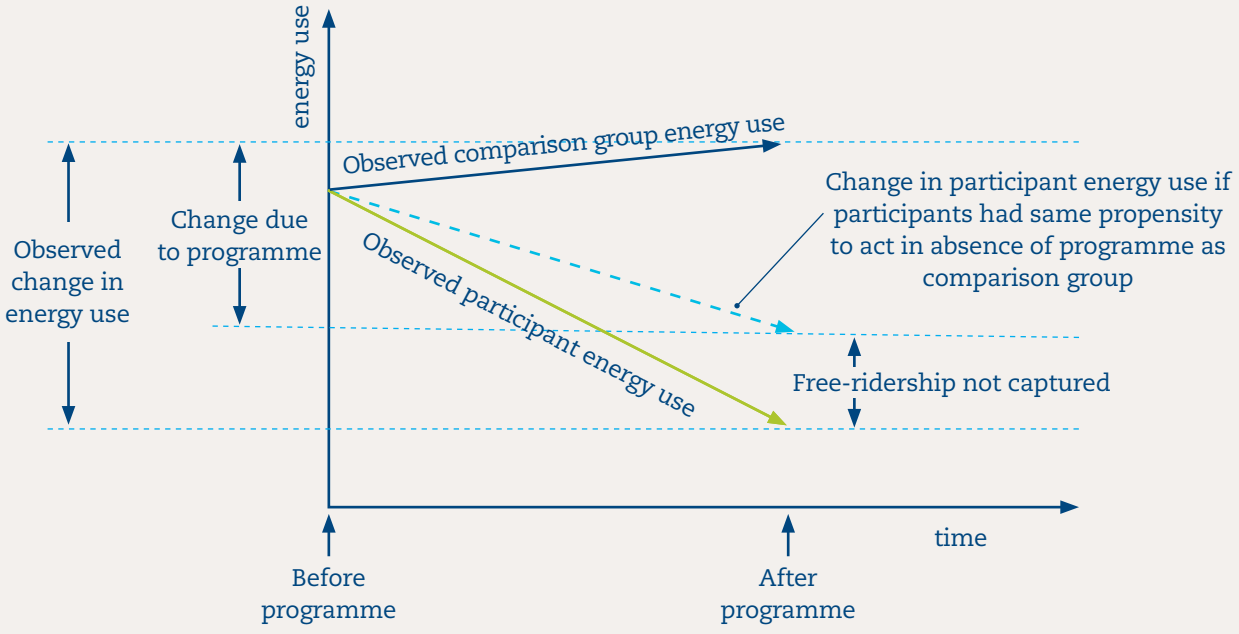
First, there may be observable characteristics that not only affect response to exogenous energy end-use determinants but also influence the decision on whether or not to participate in the programme. For example, 'households with higher initial levels of energy usage could have more potential to save, as well as more incentive to seek energy savings programs, yielding both higher savings and higher percentage of savings even in the absence of a program effect' (Johnson, 1983). Evaluators may be able to mitigate this problem if they are able to match households in participant and control groups on the basis of energy use histories, although the necessary data may be very difficult to obtain.

However, the second issue is that there are also unobservable characteristics that may impact on energy end-use and it is difficult, if not impossible, to ensure that these are very similar between participant and control groups if some of these characteristics also affect the decision to participate in an energy efficiency programme. For example, people with high environmental awareness or with high price responsiveness may be more likely than others to take part in programmes, and they may also react differently to exogenous influences. Even matching energy use histories may not fully remove this problem.

This self-selection bias introduces two types of problem for evaluation: first, the extent to which the observed differences between participants and non-participants represent an accurate estimate of the programme's effect; second, the extent to which the results of the evaluation can be generalised to other apparently similar programmes.

Frondel and Schmidt (2005) note that the self-selection effect can include higher free-ridership than there would be amongst the population as a whole. If the propensity to make the changes without the programme is the same between the participant and comparison groups, this does not affect the estimation of programme impact. However, if the propensity to make changes without the programme is actually higher amongst those that participated than amongst the comparison group, the effect will not be cancelled out by the comparison of the two groups and hence the evaluation may overestimate the true impact of the programme. Figure 2.2 illustrates this problem of 'residual' free-ridership.

Figure 2.2: the effect of free-ridership



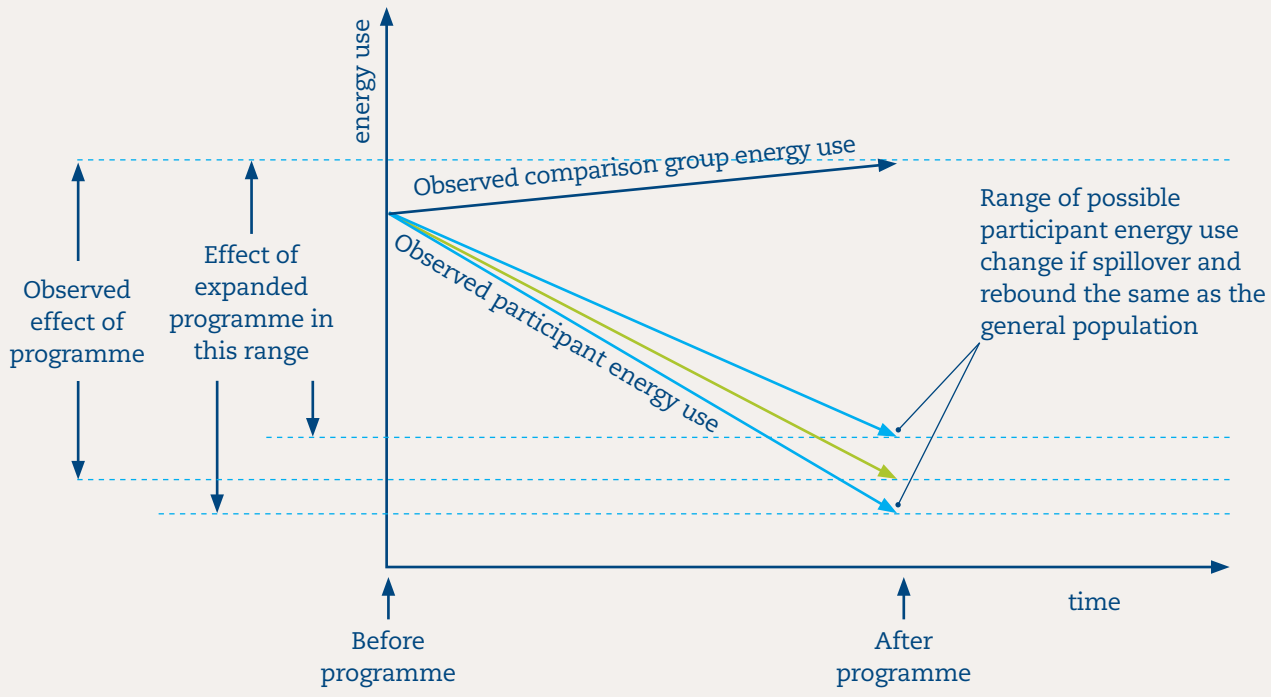
25 The comparison may be between participant and non-participant energy use after programme implementation, or it may be between the 'before to after' change in participant energy use and the equivalent change in non-participant energy use. This is discussed further in the evaluation methods section.

Similarly, the participant group may have a different reaction than the comparison group to one or more exogenous influences, such as energy prices, and this may mean that the comparison between the groups produces an inaccurate estimate of programme effect. These issues both affect the evaluation of the programme itself and the extent to which the results can be applied more generally, to other programmes.

Self-selection bias may also result in differing levels of participant spillover and rebound than the average that would occur if all households participated in the programme (Frondel and Schmidt, 2005). For example, if the unobservable characteristics include higher environmental awareness amongst participants, spillover may be increased (since they are pre-disposed to take action) and rebound may be lower than expected (if they are consciously trying to reduce energy use). If the evaluation is of a large scale programme and its purpose is simply to record the effect of that programme, neither of these elements is a problem: higher participant spillover and lower rebound are simply elements of the programme's effect and the measured difference between participants and non-participants should, and can, include these effects. However, if the evaluation is to be used to estimate the outcome of other programmes and these programmes expect to reach a broader section of the population, the outcome of these other programmes may include different levels of spillover and rebound and hence not be accurately estimated from the programme evaluation results.

Figure 2.3 illustrates the impact of differing levels of participant spillover and rebound on the generalisability of evaluation results. For simplicity, this illustration assumes that the difference between observed participant and comparison group energy use is a true reflection of the programme's effect (i.e. there is no problem of free-ridership).

Figure 2.3: spillover, rebound and generalisability



Note that for some effects, such as free-ridership, there has been significant work to estimate the magnitude of the effect. However, for others, such as the impact of self-selection on spillover and rebound, little is known.

Determining what is being measured by the evaluation

The preceding discussion assumes that the evaluation will measure whole household energy use, rather than tracking one appliance or one energy end-use. Although this is the way to gain the most accurate picture of the overall effect of a programme on a household's energy use, it may not be the most appropriate evaluation strategy for a particular programme.

For example, a programme designed on the basis of engineering estimates of its likely effect on a particular end-use may include evaluations designed to assess the accuracy of these estimates: in this case the change in the end-use in particular is the quantity of interest, and measuring changes in overall household energy use will not provide the required data.

Equally, evaluation of programmes that target individual energy end-uses may need to focus on those end-uses alone if changes in overall energy use (resulting in part from factors exogenous to the programme) may be much larger than the programme impacts themselves: attempting to separate relatively small programme effects from relatively much larger exogenous changes is unlikely to lead to accurate estimates.

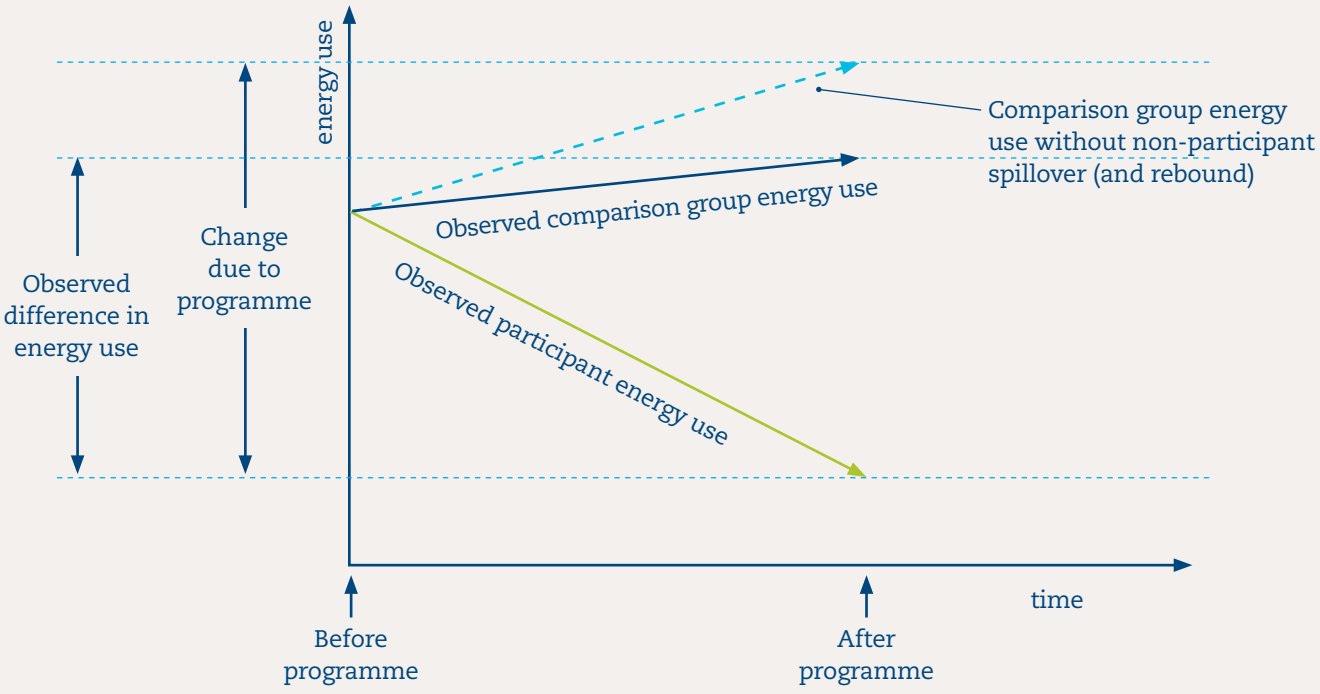
Determining who is affected by the programme

The preceding discussion also assumes that participant households are the only ones affected by the programme, as this is the focus of this study. However, a programme may result in non-participant spillover effects, if the energy use of households that do not explicitly participate in the programme is in some way affected by the programme. For example, some households may simply choose to implement changes suggested by the programme without officially participating in it.

Additionally, the programme may have an impact on the wider market for energy efficient technologies, for example by persuading manufacturers to offer more efficient models or to promote them more heavily, and by leading to increased scale of production which in turn reduces the unit costs of technologies, and this will affect both participant and non-participant energy use (Vreuls, 2005).

Evaluations based on observing changes in participant and control group energy use will capture participant spillover, but will not account for non-participant spillover (and any resulting rebound) and will hence underestimate the effect of the programme on energy use. Therefore, this study has included analysis of whether or not non-participant spillover is accounted for in evaluations where

Figure 2.4: non-participant spillover and programme outcome estimates



comparison groups are used. Figure 2.4 illustrates this effect, again assuming no free-ridership problem. As the focus of programme activity moves away from support for investment in well-established technologies and towards market transformation and changing energy-related social norms, it becomes even more important to understand the issues of non-participant spillover and indirect rebound, and the magnitude of these effects (Vine, 2013, Vine et al., 2012, Druckman et al., 2011). Approaches to capturing these wider market effects are discussed in Chapter 4.

Evaluation timeframes

The final element of the evaluation problem is how the persistence of energy end-use reductions is determined: over time, households may remove measures or revert to pre-programme energy-use actions. In addition, the technical efficiency of some measures may degrade over time. Evaluations may capture some of this effect, but are unlikely to fully reflect it as they are unlikely to continue over the full lifetime of the programme's measures. Evaluation practice in the US makes use of the concept of an 'Effective Useful Life' (EUL) to represent the combination of persistence and degradation (see, for example, (CPUC, 2006) when estimating future savings from implemented measures. Adjusting estimates for persistence will be more important for some measures and actions than others: for example, wall insulation is likely to stay in the wall for many years unless significant refurbishment work is undertaken, whereas energy efficient light bulbs may be easily removed at any point after installation.

Linked to this issue is the decision between either reporting lifetime savings from programme-induced changes or reporting only the savings in the first year after the programme. Reporting lifetime savings requires an estimate of the persistence of the changes and, where the timing of savings is important (for example, in cost-benefit analyses), a decision on whether or not future savings should be discounted and, if so, at what rate.

2.2. Evaluation methods

A number of different methods are commonly used for ex-post energy efficiency programme outcome evaluation. These can be grouped into four broad categories: engineering estimates, before-after comparisons, quasi-experimental methods and experiments. Each of these is described here, together with its potential to accurately estimate programme outcomes and the data needed for its robust implementation. Note that this discussion is restricted to methods to assess the short-term programme effects on participant households, since this is the main focus of this study. Methods to assess the effects of programmes at the wider scale and over longer timeframes are discussed briefly in Chapter 4. Note also that this section describes what the methods may deliver if implemented optimally: the practicalities of real-world evaluations are discussed in Section 2.4.

Engineering estimates

Simple engineering estimates are based on the number of measures installed as a result of a programme, the efficiencies of these measures, the efficiencies of the technologies they replace, and observed or estimated hours of use of the technology, before and after the programme. The estimated hours of use are frequently based on prior measurements, either in previous programmes or in laboratory tests that simulate in-home use of the technology. Efficiencies may be those declared by the equipment manufacturer, or may also be derived from test results. More complex engineering approaches use site visits and measured technology-specific data (e.g. using data loggers that record when a piece of equipment is operating) to include more accurate representation of use patterns (Cabrera et al., 2012).

At their simplest level, these approaches do not address many of the key elements of the evaluation problem. Free-ridership is not accounted for and, when calculations are based on estimated hours of use, the following inaccuracies may also arise:

- Exogenous influences (for example, changes in energy prices) between the time the estimated use patterns were developed and the time of programme implementation will not be taken into account, and these may have changed the way that the equipment concerned is used (higher energy prices may reduce the hours of use of a heating system, for example);
- Participant spillover will not be accounted for[26];
- Rebound effects will not be accounted for;
- The estimated use patterns may have been developed for an 'average' household; those taking part in the programme may have very different use patterns (either because the programme focuses on a particular type of household or resulting from self-selection to participate in the programme).

26 The element of participant spillover relating to the measures supported by the programme may be captured if evaluators note any additional installations of these measures whilst surveying households to verify programme data.

Where estimated use patterns are based on modelled buildings, rather than measured use patterns (for example, using the outputs from building energy rating systems) these may be inaccurate, even for the 'average' household. There has been limited work to date on the extent of this inaccuracy. However, Sunikka-Blank and Galvin (2012) introduce the idea of 'prebound': the difference between energy use in a home as predicted by a building energy rating system and actual energy use, before implementation of a programme. Based on primary data from Germany, and studies in the UK, France and Belgium, the authors suggest that actual home energy use is, on average, around 30% lower than the rating predicts[27]. They also note that difference is greater for homes with a low energy efficiency and decreases as predicted energy efficiency increases until a level of theoretical efficiency is reached beyond which actual energy use is greater than predicted. The implication of this effect is that evaluations based on simple engineering estimates of before and after technical efficiencies and modelled use patterns will overestimate the effect of the programme.

When 'before and after' observations of equipment use are employed the issue of inaccuracy in estimates of pre-programme energy use is avoided; rebound or spillover relating to the use of the equipment in question will be captured, and the difference between participants and 'average' households is not relevant. However, the following potential sources of inaccuracy remain:

- Any effect of exogenous factors on the use of the equipment between the 'before' and 'after' observations will be attributed to the programme;
- Participant spillover relating to other household energy uses will not be reflected in these estimates;
- This approach will capture rebound in use of the equipment in question, but not within broader home energy use.

Adjusting for exogenous influences, rebound, spillover and free-ridership

Although basic engineering calculations do not take into account many aspects of the evaluation problem and hence could produce inaccurate estimates of programme outcomes, enhanced algorithms may be adjusted to take into account factors such as rebound, free-ridership or spillover (SRC, 2001). Some exogenous influences, such as changes in weather, are routinely accounted for using adjustments to the algorithms used, using weather correction or weather normalisation[28].

Adjustments may be made using factors based on prior experience for a given measure or programme delivery mechanism (for example, in the evaluation of the UK's Energy Efficiency Commitment programme, as reported in Rosenow and Galvin (2013)) or evaluators may combine an engineering calculation with self-reporting of programme influence to adjust for free-ridership, (for example, as defined in the California Public Utility Commission's evaluation protocols for programmes requiring basic[29] evaluations only (CPUC, 2006)).

Data requirements

The biggest benefit to engineering estimates is that they may require relatively few data and hence are cheaper and easier to implement than the quasi-experimental or experimental approaches described below. Note however, that the challenges of installing and retrieving data loggers or sub-meters to measure usage patterns are non-trivial and should not be underestimated. Engineering estimates, adjusted as described above, may provide an acceptable level of accuracy for programmes implementing well-understood measures that have little impact either on householder actions or on the wider market for energy efficiency measures. These methods can also be useful when exploring the interaction between different technologies and, in general, as a reasonableness check on the results from other methods (SRC, 2001).

Before-after comparisons for participant households

In this approach[30], the counterfactual is defined as the energy use of participant households before the programme. Therefore the comparison group is the participant households themselves. Assuming that total household energy use is monitored, participant spillover and rebound effects will be captured (although not separately quantified) by this method.

Sources of inaccuracy inherent in the simplest forms of this approach are:

- The assumption that there are no exogenous influences (e.g. income or price changes) acting at the same time as the programme that may alter participant household energy use;
- Free-ridership, which is not accounted for.

27 The models on which these rating are based make assumptions about, for example, standard internal temperatures and hours of heating. If these are not achieved in practice, the outputs from the model will not offer an accurate estimate of the building's energy use.

28 Weather correction adjusts changes in energy use to account for the difference in temperatures between 'before programme' and 'after programme' measurement periods; weather normalisation adjusts energy savings estimates so that they relate to 'average' weather conditions rather than the specific conditions at the time of programme implementation.

29 The protocols define a number of different levels of rigour for evaluations, depending on what is already known about the programme's impact and/or what else needs to be known. 'Basic' is the lowest level of rigour required.

30 Sometimes referred to as billing analysis, together with the quasi-experimental approaches detailed below.

Adjusting for exogenous influences and free-ridership

As with engineering estimates, simple before-after estimates can be enhanced to adjust for some exogenous influences and free-ridership. Weather correction or normalisation is routinely carried out as part of this type of analysis, and corrections for free-ridership may be made using the same techniques applied to engineering estimates.

Data requirements

Before-after comparisons are only appropriate if reliable data is available on energy use for the periods before and after the programme is implemented. To ensure that seasonal variations in energy use do not bias the results, at least 12 months billing data are required for each of the before and after periods (CPUC, 2006).

Quasi-experimental approaches

These are a group of methods that may be applied when evaluators would ideally like to follow an experimental approach (see below) but are not able to because of constraints on data availability or on the way the programme is implemented (Vine et al., 2014). They range from simple comparisons between post-programme energy use of participants and a comparison group, to multivariate analyses that offer the most complex but potentially most accurate outcome estimates[31]. All rely on measured data and all involve comparisons between programme participants and a group of non-participants. The type of comparison differs between the different methods, each of which is described below.

Frondel and Schmidt (2001) suggest that these quasi-experimental approaches, if implemented robustly, can be 'powerful competitors to experimental studies'. Referring to econometric evaluations of environmental policies in general, Greenstone and Gayer (2009) note that, in these approaches, whether a household is in the group of participants or the comparison group is not randomly determined (as in an experimental approach, see below), rather it is determined by 'nature, politics, an accident, or some other action beyond the researcher's control'. They contend that it may still be possible to produce valid estimates of the effect of a programme on participants if assignment to one or other of the groups is not related to the determinants of outcomes (i.e. if there is not a significant effect of self-selection bias).

The accuracy of the estimates produced depends in part on the limitations of the method chosen and in part on the detail of its implementation, but these methods do all have the potential to produce robust and useful evaluation results. Self-selection bias does however remain an issue in many of the approaches. The more complex methods in this group use regression analysis to model the effect of multiple influences on the energy use in participant and comparison group households. These models will only produce accurate estimates if they capture the effects of all the key factors that are have a significant impact on household energy use, and if the functional form of each variable is specified correctly (Greenstone and Gayer, 2009, CPUC, 2006). Robust evaluations using these methods should consider a range of model specifications and test each for robustness (TecMarket Works, 2004).

Matching participants and comparison group households

There are a number of different ways in which evaluators can match participants with comparison group households, and these are described briefly here. At a basic level, a comparison group can be formed from a pool of households with similar key characteristics thought to influence energy use, such as income level, housing type, household size, or geographic location. This option is the most straightforward and requires less data and less data processing; hence it is likely to be the cheapest option available. However, a comparison between two groups matched in this way is not likely to overcome selection bias issues particularly well.

Exact matching

To increase potential accuracy, small groups within the overall sample can be matched with small groups of non-participants from the comparison group sample, based on similarities in the key characteristics mentioned above and in energy use history (level of energy use over the 12 months preceding the programme and, ideally, also patterns of variation in energy use over this period). The aim here is for the closer matching of participant and non-participant households to ensure that the impact of exogenous variables is similar across participants and non-participants in matched groups, and hence for the issue of self-selection to be dealt with Johnson (1983).

Participants and comparison group households can be matched on the basis of 'propensity scores'. In this method, regression analysis is used to estimate the probability that a given type of household will participate in a programme, based on the observable variables mentioned above. Participants and comparison group households are then matched on the basis of similarity in their probability of participation. Alternatively, households can be matched using the Euclidean Distance between them. This type of matching involves calculating the square root of the summed squares of the differences between each of the variables being used to match the households. For both propensity score matching and Euclidean distance matching, the sample of households may first be stratified, for example to group together all

31 However, the accuracy of the estimates relies on gathering sometimes very large amounts of data. As this is not always available, the theoretically possible accuracy may not always be delivered in practice.

households on a specific low-income tariff or in a given geographical area (Hanna and Marvin, 2013). Choice between the two methods will depend on the available data: for example Hannah and Marvin (ibid.) suggest that propensity score matching is more effective for smaller participant and comparison group pools, but may not produce optimal matches because of the small group sizes, whilst Euclidean distance matching is more effective for larger groups although the data processing necessary requires significant computer processing power and time.

Comparison methods

There are two main types of comparison used in quasi-experimental (and experimental) approaches: cross-sectional and difference-in-differences. In a cross-sectional approach energy use in participant and comparison group households after the implementation of the programme is compared. Difference-in-differences is a combination of before-after and cross section approaches, comparing the change in participant households' energy use over the programme period with the change in non-participant households' energy use.

Either approach can be used to compare the whole participant group with the whole comparison group or, where exact matching techniques have been used, to compare each matched set of households, after which a (weighted) average of the results is used to estimate the overall programme outcome.

The extent to which either approach accurately estimates programme outcomes depends on the extent of self-selection bias remaining after the comparison group has been matched with the participant group. Although careful matching on pre-programme energy use should reduce this bias, it is not possible for evaluators to be certain that it has been removed entirely (Provencher et al., 2013). If some selection bias remains, the estimate will not accurately account for differences in the effect of exogenous variables on energy use or the extent of free-ridership within the participant group. As with previous methods, correction factors can be applied to mitigate these problems.

In addition, quasi-experimental methods assume that the programme has not had any effect on households in the comparison group. If there has been non-participant spillover within the comparison group, this could result in an under-estimate of programme effects (as illustrated in Figure 2.4).

Data requirements

Quasi-experimental methods require more data than engineering approaches or simple before-after comparisons, particularly when exact matching methods are used to reduce selection-bias issues. Achieving close enough matching of sufficient groups of participant and control households requires data on significant numbers of households and also time series data for at least 12 months prior to programme implementation (to demonstrate close matching of energy use patterns). Hanna and Marvin (2013) suggest that at least four times as many comparison group households are needed as participant group households to enable good matches to be found.

The size of the sample of households needed to produce results that are considered to be statistically robust[32] depends on the likely size of the energy saving effect produced by the programme (see for example (HM Treasury, 2011), box 9E). As an illustration, for energy savings in the region of 10 per cent of the total energy use, a sample of 5-600 households will be sufficient whereas for energy savings in the region of 1-2% of total energy use, a sample of 12-15,000 households is needed (Vine et al., 2014).

Experiments

Experimental approaches, specifically Randomised Control Trials (RCT), are considered to be potentially the most accurate way to estimate programme outcomes. In an experimental evaluation design, households are invited to participate in a programme and those that agree are randomly assigned to participant and control groups. Since both groups of households volunteered to participate, the issues of self-selection bias should be avoided (Frondel and Schmidt, 2001).

Participant spillover, rebound, and free-ridership issues are all accounted for in the estimates produced by this approach. Careful experimental design may ensure that there is no non-participant spillover[33] but this will therefore result in a potential difference between the trial and the full-scale programme and hence could be a source of inaccuracy.

One further potential issue with this approach is that the process of randomisation may disrupt the status quo, making participants in the treatment group particularly energy conscious and hence leading to an overestimation of the programme's true effects in the absence of the experimental (Davis et al., 2013)[34].

32 Robust here means accurate and precise, and statistically significant.

33 If it is possible to define a control group that has no contact with the participant group and an intervention that has no effects outside the participating households.

34 This effect is sometimes referred to as the 'Hawthorne effect', after experiments in the 1920s looking to increase worker productivity at the Hawthorne works outside Chicago. Worker productivity seemed to increase in the experiments regardless of what changes to the environment or working practices were made, leading to the idea that the change was because of the additional attention being received by the workers during the experiment. However, later analysis of the experimental data suggested that the observed effect may have been due to timing rather than either the experimental changes or the fact of being experimented on. The fact of being observed in an experiment is still thought to have an effect on actions of participants, but the magnitude and nature of the effect is difficult to determine and depends on precise experimental conditions (MCCARNEY et al., 2007).

Data requirements

Experimental evaluation designs have very similar data requirements to quasi-experiments, with a need for time-series of pre- and post-programme energy use for appropriately sized samples of households. Here again the sample size will depend on the expected scale of programme outcome in comparison with total energy use, which for some programmes will result in very large samples being required.

2.3. Summary comparison of methods

Table 2.1 below summarises the methods described above, in terms of: the extent to which they address issues in the definition of a counterfactual; key pros and cons of each method when compared with the others; and when, in theory, they should be used.

Table 2.1: summary comparison of methods

Method	Issues in defining the counterfactual: exogenous influences	participant spillover	rebound	self-selection bias	free-ridership	non-participant spillover	Key benefits	Key drawbacks	When to use
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 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 there is not likely to be a large non-participant spillover effect
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; where non-participant spillover is not a major issue
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

2.4. Evaluation in practice

It is possible to argue from a theoretical point of view that the level of accuracy of evaluation results will increase as one progresses down Table 2.1 from engineering approaches to Randomised Control Trials. However, as Table 2.1 shows, each method has drawbacks as well as benefits, and the most appropriate approach in practice may depend on a number of different constraints or opportunities offered by the particular programme being evaluated. It is also worth noting that the cost of an evaluation is likely to increase as one progresses from simple engineering estimates to Randomised Control Trials.

This section discusses some of the constraints evaluators face in practice and then summarises what this means for the development of an appropriate evaluation strategy for any given programme. The implications of the issues discussed for the review of evidence in this study are described at the end of this section.

Data issues

Whatever evaluation approach is taken, the accuracy of the results will depend on the quality of the data collection and manipulation. Key elements of high quality data (Vreuls, 2005) include:

- Accuracy;
- Verifiability;
- Lack of bias, and
- Availability over the necessary time period.

Accuracy of energy use data can be a particular issue for evaluators in the UK, as household energy bills are often based on estimated usage data rather than meter readings. This may be less of an issue in other countries, such as the US, where monthly meter reading is more usual. Smart meter introduction offers a potential solution to this problem.

Self-reported data (e.g. from surveys of households for use in estimation of free-ridership levels) can also pose issues linked to accuracy: whilst Vreuls suggests that 'self-reports can be very useful in characterising programme effects on key householder decision elements', the extent to which respondents are able and willing to give accurate answers to questions about energy-related actions in the home, equipment purchase decisions, and the factors that affect these has been questioned (see, for example, (Peters and McRae, 2008)) although evaluation guidelines and protocols (SRC, 2001, CPUC, 2006) suggest questionnaire design and implementation methods to mitigate some of the problems and increase the reliability of the results.

Verification of data should be a key element of robust evaluations, and indeed evaluation protocols and guidance refer to elements such as verifying that recorded measures have actually been installed / installed correctly (SRC, 2001, CPUC, 2006). However, budget constraints may mean that only a small proportion of recorded installations are checked, or that evaluators rely solely on programme records for some elements of their data.

In all evaluations, information on pre and post intervention energy use needs to be based on a sufficiently large sample to produce meaningful results that accurately reflect the full programme (i.e. that avoid bias). Sampling at an appropriate level[35] may be difficult, depending on the willingness of participants and non-participants to provide information to the evaluation team. As noted by Vine (2013), finding an appropriate control group for an evaluation, which mitigates selection bias problems whilst at the same time not introducing errors from unacknowledged non-participant spillover, may be increasingly difficult when an increasing proportion of the population is in one way or another affected by energy efficiency programme activities. Evaluation aims linked to the effects of a programme on different sub-groups within the overall participant group increase the volume and complexity of data required, since samples for each sub-group need to be large enough to produce significant results and also designed to avoid bias.

Ensuring that the required data are available across the whole time period required can be a challenge. This issue can be exacerbated if the design of the evaluation is not part of the initial planning of the programme and hence if pre-programme energy-use information is not collected in sufficient quantity and quality.

Ensuring high quality data also involves careful checking of data before its use but little attention is given to this data cleaning and manipulation in guidance documents. Neither the IEA nor the EC guidance includes any detail on this element of evaluation practice. However, the California Public Utilities Commission Protocols (CPUC, 2006) note the potential impact of these elements on the evaluation outcome and stress that methods used must be clearly reported.

Implementing Randomised Control Trials

RCT may be seen as the 'gold standard' for outcome evaluations, but they are rarely used in practice. There are a number of practical barriers to their use that at least partially explain this.

Programme administrator unfamiliarity with experimental approaches, together with balancing competing objectives during programme design, can lead to refusal to fund experimental evaluations and/or programme design that does not permit the necessary data collection. They may also wish to implement a full-scale programme, with rapid evaluation of annual results, rather than first wait for an experimental evaluation of a pilot (Vine et al., 2014).

35 See the 'data requirements' discussions in Section 2.2 for more on this.

Experimental approaches can be very difficult to implement if they raise ethical concerns. For example, regulators or government may require that interventions are available to everyone. This is particularly an issue for programmes that aim to alleviate fuel poverty. A number of experimental studies of the health impacts of fuel poverty programmes have successfully dealt with ethical issues, for example by providing measures to the control group as soon as the experiment is complete (Osman et al., 2008) . Formally, this approach is referred to as 'RCT with delayed treatment' and may be considered the 'next best' approach to pure RCT (Vine et al., 2014). However, even if ethical concerns are addressed, it remains difficult in this type of project to ensure that researchers are blind to which group (intervention or control) households are in and hence are not tempted to offer alternative help to control group households who may be in severe need.

In addition, the ability to control conditions closely enough to perform a robust experiment is rare when the experiment concerns a change to a system as complex as the use of energy in a home: experimental approaches are more feasible when there is a direct and simple relationship between the programme and the evaluated outcome; when the outcome is large in relation to other 'background' changes in the measured variables, and when the effect is realised within a short time period (HM Treasury, 2011).

A pragmatic response to these issues in many cases is the use of quasi-experimental methods.

Transferability of evaluation findings

There is increasing recognition that programme evaluation needs to offer more than a check on the effectiveness of a particular programme: it needs to deliver useful information for designers of future initiatives. A first requirement for this is that the outputs from an evaluation can be fully understood by others.

Most evaluations are carried out to inform programme implementers about the outcomes of their programme(s). Vine et al (2012) note the lack in the US of national protocols that evaluators are required to follow, and a similar situation exists in the UK and EU. Guidance manuals, for example (SRC, 2001); (Vreuls, 2005); (HM Treasury, 2011) or (Schiller, 2007), may increase standardisation in evaluation practice, and the need for international comparability of outcomes is increasingly recognised[36]. However, there are many remaining inconsistencies between studies. For example, Vine et al note the differences in the definition of net savings between different utility regulators in the U.S., and Broc et al (2011) point out differing approaches to timeframes considered (annual versus discounted lifetime savings) and the treatment of free-ridership and spillover between French national systems and the reporting requirements for EU-required National Energy Efficiency Action Plans. Unless these elements of each evaluation are fully explained, it is difficult to apply the results of one evaluation to the development of a different programme.

A second requirement for transfer of lessons is that programme designers understand the detail of the design and operation of previous programmes and the context within which they were implemented, so that they can understand why a programme had the effect that it had. As part of this, it is important to understand the effect that any selection bias will have on the application of evaluation results: the results for any given evaluation may not be generalisable to a larger scale programme or a similar programme in a different location. For example, Davis et al (2013) reviewed 32 different pilots of various types of smart metering programme in the US and Canada and they noted that volunteers for dynamic pricing or in-home display trials tended to have relatively high income and education levels and hence the results of the trials may not be applicable to roll-out of the measures to the population as a whole.

Vine et al (2014) note this difference between the participants in an experiment and the general population as one of the main threats to generalisability of evaluation results, and also note the potential effect of a change in setting between an experiment and its generalisation. They point out that the requirements of experimental design can mean that RCT based evaluation results can be particularly difficult to generalise.

Whilst RCT may be the optimal method in theory to determine the size of the outcomes from a programme and quasi-experimental quantitative methods are a good alternative, it is very difficult to examine why people have reacted to a programme in the way they have using these methods. Some information can be gleaned from quasi-experimental methods applied to segmented participant groups, but the data required to allow statistically significant findings for multiple groups can be prohibitively numerous and difficult to collect. For formative evaluations, where the results will be used to inform scaling programmes to higher levels of uptake, the use of qualitative methods to explore why people participated and achieved energy savings can be as important, if not more so, than quantitative methods to determine relatively precise estimates of average or total savings.

36 See, for example, the World Resources Institute Greenhouse Gas Protocol work on mitigation accounting: http://www.ghgprotocol.org/mitigation-accounting

Developing an appropriate evaluation strategy

In practice, the following contextual factors are important in shaping the strategy for evaluating any given programme: the programme's objectives; the current stage in the programme's lifecycle; the policy or regulatory framework and how the evaluation results will be used; and the available budget or relative priority placed on comprehensiveness and accuracy by the organisation commissioning the evaluation (Vreuls, 2005).

In response to these contextual factors, the evaluation approach may be comprehensive (using a large data sample and accurate estimation methods), targeted (using a smaller sample and employing less data-intensive methods that produce reasonable estimates) or simply a review (using engineering estimates combined with basic programme delivery statistics).

Vreuls goes on to suggest that the following factors imply a need for more comprehensive evaluation: a large funding outlay; large expected savings; promotion of a new technology or a new delivery method; the potential for significant expansion of a pilot or changes to an established programme.

Guidance produced on behalf of the European Commission (SRC, 2001) suggests that comprehensive evaluation may be appropriate for pilots or for programmes in their second or third year of implementation, whilst more mature programmes may need a lighter touch. The guidance also notes that full implementation of an in-depth analysis of cause-and-effect relationships is not going to be cost-effective for most (smaller) programmes and for these a simple engineering approach to outcome evaluation may be sufficient if there is confidence in existing understanding of the effects of the measures and actions supported by the programme.

It is difficult to define what is meant by 'cost-effective' evaluation. The resources devoted to an evaluation are rarely mentioned in the peer-reviewed literature, and although commissioners of evaluations may work to 'rules of thumb' about approximate proportions of programme budgets to spend on evaluation, these are rarely published or discussed. UK Treasury guidelines (HM Treasury, 2011) to government programme administrators offer no suggestion of the level of financial resource to be used for evaluations, but rather a series of factors that need to be taken into account when determining this, echoing many of the factors already mentioned here. Table 2.2 summarises the key points in Table 4C of this guidance.

Over time, and at a policy level, the use of mixed evaluation methods is likely to be the most robust strategy. For example, alternating between engineering estimates and quasi-experimental methods to evaluate successive years of a multi-year programme will provide some sense-checking of the results of each method. Equally, the results of evaluations of a series of individual programmes, carried out using one or more of the methods described above, may be compared with result from one of the top-down methods of examining the overall outcome of energy efficiency action at the State or national level (see Section 4.9 for more on this alternative approach).

Table 2.2: factors to consider when resourcing an evaluation

Factor	Reason
Innovation and Risk	High risk or innovative policies need robust evidence to show whether or not they are working as expected
Scale, value and profile	Programmes that are large or high profile need robust evaluation to meet accountability requirements
Pilots	Evaluation needs to inform future activities
Generalisability	If there is the potential for the results to be more widely relevant, then the evaluation needs to be robust enough to provide confidence in this generalisation
Influence	Greater resources may be justified if an evaluation may report at a strategic point in time or if it will fill an important evidence gap
Variability of impact	Uncertain outcomes or behavioural effects that are more difficult to isolate may require more extensive evaluation
Evidence base	Evaluation is likely to require more resources where the existing evidence base is poor

Implications for this study

The importance of practical considerations and constraints in the choice of an appropriate evaluation strategy has implications for how this study can assess the robustness of the evidence base on energy efficiency programme outcomes. A series of key questions will be asked in relation to each piece of evidence (see the framework for assessing the evidence base, below). These aim to determine whether the evaluation strategy uses the best available methods given the specific circumstances in which the evaluation is being carried out and whether the methods that are used are implemented well, with their limitations acknowledged and, where possible, adjusted for.

Some of the evaluation methods described above are more commonly used than others, as the desire for accuracy is balanced by data and time constraints. Also, certain types of evaluation are more prevalent for certain types of programme. Since each method has its own strengths and weaknesses, any concentration on one or more methods may result in a systematic bias in

the results. This study aims to identify any such bias, if it exists, and comment on its possible importance.

The transferability of evaluation results is a key issue for the study: are the results defined in the evidence base transferable to the current situation in the UK? The study will look at the range of outcomes quantified in the literature and examine the extent to which these quantifications are accompanied by sufficient detail on programme design, implementation and context to enable a judgement on the likely position within that range for the outcomes of a programme to be designed today.

2.5. Framework for assessing the literature

Based on this review of evaluation issues, methods and practicalities, the following framework was defined for the assessment of the evidence base. Firstly, for each paper within the evidence base:

1. Characterise the programmes evaluated to facilitate comparison between different studies.
2. Explore the methods used and the quality of their implementation; and note the results obtained.

Following this, for the evidence base as a whole:

3. Review any systematic bias in the evidence and explore how important this may be.
4. Summarise quantified outcomes and qualitative evaluation results; discuss what degree of confidence can be placed in these results and the extent to which they are transferable.
5. Discuss where the focus of evaluation effort should be in the future.

Characterising programmes

The programmes evaluated were characterised so that results for similar types of programme can be collated and discussed. Information was recorded about:

- programme type;
- the location and scale of the programme;
- who implemented it;
- what key supporting policies and programmes were in place at the time of implementation;
- what types of housing and households the programme was aimed at, and
- who commissioned and who carried out the evaluation.

Methods used and the quality of their implementation

Papers were reviewed to determine whether or not they contain sufficient information for the evaluation methods used, and the quality of their implementation, to be determined. If there was sufficient information, the following key questions were asked of each piece of evidence:

- What evaluation methods are used?
- Does the evaluation demonstrate an understanding of how the programme is likely to affect energy use, and hence seek to collect and use appropriate data?
- Is the scale and nature of the evaluation appropriate for the programme size and stage, and level of existing knowledge about outcomes?
- Is the choice of evaluation method appropriate for the available data?
- Are the limitations of the evaluation acknowledged and, where possible, adjusted for?

Evaluation methods were noted for use in the discussion of potential bias. Each paper was then given a numerical rating between 1 and 4 for each of the remaining questions, based on the extent to which it was answered positively. An overall average numerical rating was also given, based on these individual ratings, and this was used as an indicator of the overall usefulness of the piece of evidence in the discussion to follow. A copy of the matrix used in the assessment is provided in Appendix C. Assignment of the ratings against each question involved a judgement of elements of the study, based on the information in the published paper. The factors taken into account in this judgement are explained below.

Understanding how the programme is likely to affect energy use

Ideally, an evaluation would start from the point of a clearly defined theory of how the programme was expected to affect energy use, for example through the use of logic models[37] for the programme (Vreuls, 2005). Some of the literature does indeed explicitly start from this point, although in many cases papers do not state the theories that they are testing. In the latter case, a judgement on the level of understanding was based on this author's knowledge of how the programme is likely to have affected energy use and the extent to which the evaluation reflects that.

37 Logic models are representations of how a programme is thought to deliver its outcomes. For example, given inputs lead to a set of programme activities; these in turn produce outputs, which then lead to a series of outcomes. The routes through which each of these cause and effect links work are defined in the logic model of the programme.

Designing an evaluation based on the logic of how a programme is expected to have an effect is clearly necessary when the evaluator is interesting in how the programme is delivering its results. However, this study assumes that this approach is also necessary when an evaluator is simply interested in what the effect of the programme is: if the evaluator does not understand the programme logic, they may not choose the most appropriate data to measure its outcomes.

This question was also used to reflect on the extent to which the evaluation could capture unexpected effects (i.e. were the data collected sufficiently broad to uncover effects that were not part of the programme theory; for example participant spillover from a programme targeting electricity use that resulted in reductions in gas use).

Appropriateness of the scale and nature of the evaluation

As discussed earlier in this chapter, the necessary comprehensiveness of an evaluation will depend on the nature and stage of a programme. This question considered whether the scale of the evaluation was consistent with these aspects of the programme. Note that this had to be judged qualitatively, as there is very little information about costs of evaluations in the literature, as discussed previously.

It also considered whether the evaluation method(s) chosen were consistent with the existing level of knowledge about programme effects (for example, use of an enhanced engineering approach when the effects were thought to be well understood already, or use of a Randomised Control Trial for a pilot of a completely new approach, when its effects could potentially be significant).

Coherence of evaluation method and available data

As noted previously, the quantity and quality of data that can be collected for any given programme should influence the choice of evaluation method. This question was used to check that the evaluation methods used were appropriate given the amount of data available to evaluators and their judgement on the quality of these data. For example, where quasi-experimental methods were used, was the choice of matching method based on the best available comparison group; and was the complexity of statistical analysis justified by the level of confidence that could be assigned to the results produced.

Acknowledgement of limitations

All evaluation methods have limitations, and this question checked whether evaluators acknowledged the limits of the method that they had used, together with any data issues encountered, and discussed the impact that these had on their confidence in their results. This question was also used to check whether papers suggested that corrections to the results could be used to account for the limitations.

Credit was given here to evaluations that used multiple methods and triangulated results; those that compared their results with previous evaluations of similar programmes, in particular those that had used different methods; and evaluations that commented on how their results compared with what had been expected, attempting to explain any differences.

Systematic bias in the evidence base

The study also included a review of the extent to which the evaluation methods used may be introducing a systematic bias into estimates of effectiveness, either for energy efficiency programmes overall or for particular types of programme. For example, if all utility investment programmes were evaluated using methods that ignored rebound, it may be that the effects of these programmes are systematically being overestimated. Similarly, if the non-participant spillover effects of innovative finance offerings are ignored in evaluations using comparison groups, the effects of the programmes may be systematically underestimated.

Where the potential for bias in the results is identified, the possible implications of this are discussed, although the extent of the problem created may be difficult to define. For example, self-selection is an issue in many, if not most, evaluations. However, as noted by Greenstone and Gayer (2009), it is difficult to predict either the magnitude or the direction of the impact that it may have on evaluation results.

Collating quantitative and qualitative outcomes

The results of the evaluations that are judged to be of sufficient quality have been collated into an overview of knowledge for each of the programme types defined earlier, and these are presented in Chapter 4. Key gaps where further work is needed are identified, and the implications of these for evaluation practice are presented in Chapter 5.

Initially, the quality threshold for including papers in the evidence base for this report was set at an average score of 3 or higher against the key questions asked. In addition, information has been taken from papers with an average score of 2.75 where the average was lowered by a negative answer to the final question (whether or not the paper acknowledged limits to the evaluation method used) since this seemed to be an issue for a number of otherwise high scoring papers. Additional, qualitative information has been drawn from a number of papers that did not provide sufficient detail on evaluation methods to be scored within the framework but were nonetheless deemed useful for the discussion.

Review of evaluation practice

Finally, evaluation practice as represented in the literature was reviewed to determine whether there are specific issues (e.g. treatment of self-selection bias or inclusion of non-participant spillover) that are still not commonly addressed in evaluations, and also whether or not individual programme evaluations are recording sufficient information to be useful for future programme design. The results are presented in Chapter 4 and their implications discussed in Chapter 5.

3. Overview of the evidence base

The evidence base for this study comprised peer reviewed journal and conference papers only. It was further restricted, by time constraints and also a desire to focus on evaluation practice in recent years, to journal articles published in the time period 2000 to 2013 and conference papers from the 2010 and 2012 American Council for an Energy Efficient Economy (ACEEE) summer studies, the 2009, 2011 and 2013 European Council for an Energy Efficient Economy (ECEEE) summer studies and the 2010, 2011, 2012 and 2013 International Energy Program Evaluation Conferences (IEPEC). Where evidence for a particular programme type was scarce, key papers from outside this scope have been identified and their results added to the discussion of those particular programme types. The overview presented in the remainder of this chapter concentrates on the scope as originally defined.

3.1. Where the literature was located

The literature is widely spread across a number of sources. Conference proceedings were the richest source of useful material, with 52 of the 93 papers included in the initial review coming from the eight sets of conference proceedings listed above. Nine energy-related journals held 22 papers, three building-science publications held 9, four (energy) economics journals held 6, and four environmental science / geography publications held the remaining 4. Once the initial review was completed, papers were excluded that provided little or no information on programme outcomes or where the regulatory, market and/or cultural context meant that the results could not be transferred to the UK situation. This left a core evidence base of 68 papers. These are the papers described in the remainder of this Chapter.

3.2. Types of programme evaluated

The evidence base is somewhat dominated by utility activity, which is perhaps not surprising since this has been, and remains, one of the major routes to implementing household energy efficiency programmes, at least in the UK and US[38]. Traditional utility Demand Side Management (DSM) programmes or other (e.g. government-led) large scale investment programmes were the subject of 14 papers, whilst billing feedback and time of use pricing were the focus of 17 papers.

The effects of building codes / regulations and energy performance certificates for buildings were covered in eight papers. There were also eight papers covering a range of information and advice programmes. Seven papers reviewed the effects of programmes targeting low income households.

Innovative finance mechanisms were the subject of six papers, whilst community or peer-group approaches were the focus of five papers. Both of these types of programme have increased in importance in very recent years, and hence it may be that the volume of evidence on their effectiveness may increase in the near future.

There were only five papers considering aspects of appliance market transformation programmes. This relative lack of evidence for a well-established set of programmes could be a function of the cut-off dates for the literature search: the initial set of papers retrieved did contain a number looking at appliance efficiency programmes in Asia and South America, but these were excluded because their context rendered their results non-transferable; therefore, the section on appliance market transformation will also refer to earlier evidence for similar initiatives in Europe, North America and Australasia, where this has been found.

Note that many of the programmes evaluated are actually combinations of two or more of the programme types described in Box 1.1. In general, it is clear from the evaluation report that the programme is considered (by its implementer or by the evaluator) to be primarily of one type and hence this is where it is reported here. For example, many utility programmes include elements of information and advice, as well as subsidies or direct installation of efficiency measures, but these elements are often viewed as supporting the main investment programme and hence the evaluations will be reported in the section on 'large scale investment and refurbishment programmes'. Whilst it would be preferable to report separately on the outcomes of each of the different programme elements, this level of attribution is very difficult and is usually not attempted in the evaluations.

38 Activity in other EU countries has been less dominated by utility programmes, and hence the literature may be showing a UK/US bias. However, it is difficult to judge the extent of this problem. The MURE database (http://www.measures-odyssee-mure.eu/) lists 559 'measures' aimed at households in the EU28 Member States and Norway, together with further 'general' measures including utility programmes. However, it is outside the scope of this report to review the summaries for these measures in sufficient detail to form a view on the dominance of one type or another.

3.3. Where, when and who

The evidence base was balanced between papers covering programmes in the US and Canada (34) and those covering the UK and other European countries (33). In addition there were four papers reporting programmes in Australia and New Zealand, and three reporting programmes in Asia[39].

Twenty seven papers concerned programmes implemented at a national scale; 25 at the regional or State level; and 21 at the city or local level.

The evidence base is naturally biased towards more recently implemented programmes, since it is limited in scope to papers from more recent years. However, there are ten papers that concern programmes where implementation began before the year 2000, and four of these cover programmes that commenced in the 1970s and 1980s.

Utilities were the programme implementers in 38 of the papers, and in ten of these they were working in partnership with government or private sector / not for profit organisations[40]. National government was responsible for implementing programmes covered in 17 papers, with sub-national government involved in 14 cases. Not for profit organisations were involved in programmes reported in 12 papers, but in only four of these were they solely responsible for implementation. Similarly, private sector organisations other than utilities were involved in programmes in five papers, but in only one was a private sector organisation solely responsible for delivery.

The evaluations of the programmes were commissioned and carried out by a range of types of organisation. In 32 of the studies, it was not clear who had commissioned the work; where this was made clear, the majority of studies (30) were commissioned by the implementing organisation. Academics were involved as evaluators in 27 cases and other independents in 24. The programme implementer was involved in the evaluation in 24 cases, often in collaboration with an independent evaluator. However, there were 15 instances where programme evaluations appeared only to involve the implementing organisation.

3.4. Context

There was very little information given about the context within which programmes were implemented, with only a small minority of papers mentioning other initiatives that were active at the time the programme was implemented, or changes in key variables such as energy prices. Similarly, few explained in any detail the socio-demographics of the households targeted by the programme.

3.5. Quality of the evidence base

In total, 48 of the papers in the evidence base were assessed against the scoring matrix. Most of the 20 remaining papers did not provide sufficient methodological detail for the questions to be answered, in some cases because the main focus of the paper was on a description of the programme or a process evaluation, with a minor element concerning outcomes.

The quality threshold (average score of 3 or more) was met by 32 papers. In addition, a further 8 papers had an average score of 2.75 and in each case this was a combination of an average higher than 3 for the first three questions together with a low score for the acknowledgement of study limitations[41]. These 40 papers therefore make up the main body of evidence discussed in detail in Chapter 4, with some additional information drawn from papers that could not be rated.

Looking in more detail at the questions asked to rate the literature:

- 38 of the 40 papers scored 3 or more for the extent to which they demonstrated an understanding of how the programme being evaluated was likely to affect energy use;
- 33 papers scored 3 or more for the extent to which the scale and nature of the evaluation appeared consistent with programme size and stage and the level of existing knowledge;
- 39 of the 40 scored 3 or more for the extent to which the evaluation method was coherent with the data available to the evaluator; and
- A much smaller proportion of papers (24 out of 40) scored 3 or more for the extent to which methodological or practical limitations were acknowledged and, where possible, adjusted for.

39 The number of papers here totals more than the 68 in the evidence base because some papers cover programmes in more than one of the regions listed.

40 As mentioned previously, this may indicate a bias in the peer-reviewed, English language literature that does not fairly reflect attention given to other types of action in some EU countries.

41 The remaining 8 papers received an average score of less than 2.75 and hence were excluded as confidence in their results would be low.

Many studies measured changes in one fuel only, which leads to concern about the extent to which all possible effects on household energy use are being captured. Also, there was a general lack of information about the quality of data used in evaluations (for example whether billing data was actual metered data or supplier estimates, or whether significant data cleaning had been necessary).

In summary, the quality of the evidence base overall was reasonable, but there remain issues to address: only four papers scored 4s against all of the criteria[42], and 22 papers scored less than 3 against one of the criteria, indicating significant weaknesses in this respect. The implications of this for confidence in the results for specific types of programme will be discussed as appropriate in the relevant sections of Chapter 4.

Bias

The extent to which the evaluations accounted for the various elements of the evaluation problem, and hence the extent to which results may be subject to a systematic bias, varied across the different elements.

In general, the literature dealt very well with participant spillover and direct rebound: in both cases over 30 of the 40 papers described methods that clearly seemed to take these into account. Just over half the papers (22) dealt adequately with exogenous influences, with a further 12 papers potentially dealing well with this issue although it was not possible to be fully confident about this from the information given.

Free-ridership was less well addressed: 15 of the 40 papers clearly took account of it and a further 11 may have done so, but 14 papers clearly did not. Only 7 of the 40 papers clearly addressed the issue of self-selection bias, although a further 13 may have done so. And unsurprisingly, only 2 papers clearly covered the issue of non-participant spillover, with one further paper possibly doing so.

42 These papers concerned billing feedback studies and hence their high scores reflect a particular opportunity to conduct Randomised Control Trials.

4. Findings

This chapter presents the key evidence from the peer-reviewed literature for each programme type, offering quantified estimates of savings where possible, and discussing the degree of confidence that can be assigned to these estimates. The potential for the estimates to over- or under-state the true value of direct net savings is discussed where there are sufficient quantitative results for this to be appropriate, with reference to the elements of the evaluation problem that have been taken into account in the evaluations.

4.1. Minimum efficiency standards for buildings

Several approaches to estimating the effects of building codes or regulations in a number of different countries are presented in the literature. At the most fundamental level, there is some evidence that building codes do lead to increased energy efficiency, and that they may reduce home heating energy use, but by a smaller amount than ex-ante estimates would suggest. However, the literature offers very little useful quantitative information beyond this.

Saussay et al (2012) report on an IEA review of the effects of building codes on residential space heating in Austria, Denmark, Finland, France, Germany, Poland and the UK. Their econometric model of the evolution of space heating efficiency offers a view of the trend in space heating energy efficiency over time, but does not allow year to year comparisons. The results show that there is a statistically significant increase in space heating energy efficiency over time in all countries studied, and that energy efficiency increases as the number of years since building code energy requirements were first introduced increases. The authors are confident in the data quality for most of the key variables used and, whilst they recognise that there are elements of the model that could be improved, they express confidence in this initial result.

As the efficiency requirements defined in building regulations vary between countries, perhaps the most useful type of information about the effectiveness of building standards is a comparison of the achieved savings and those that were expected. Kjaerbye et al (2011) report on an investigation of the effects of Danish building regulations on household natural gas usage. Using econometric analysis of multiple housing datasets and actual metered gas consumption over a six year period for a sample of around 37,400 single family owner-occupied homes, they conclude that tighter building energy efficiency regulations have clearly improved housing energy efficiency. They estimate that the 1998 building regulations have reduced heating energy use, compared with buildings constructed under the previous regulations, by 7%. This compares with an ex-ante estimate of a 25% reduction. The authors explore their results by applying a number of different econometric treatments to the data, demonstrating that these make little difference to the result and hence they conclude that their findings are robust in this respect. However, the study is reliant on a number of assumptions that could perhaps be challenged. For example, although they recognise that the energy efficiency of an older home may be increased after construction and do attempt to account for this (unlike other studies), they assume that the longer a household is resident in a property, the more likely they are to have made energy efficiency investments; the reasoning behind this and other such assumptions is not clearly explained. An issue with the comparison between the modelling result and the ex-ante estimate is that the authors do not explain how the latter was produced and hence it is not possible to understand how valid the comparison is.

Rogan and O Gallachoir (2011) also use metered gas usage data to compare average actual heating energy use in 6,000 homes built to the 2002 Irish building regulations with that in 5,000 homes built to the 1997 regulations. Their results suggest that the increased energy efficiency requirements in the 2002 regulations have reduced gas usage by 10.1% relative to the requirements in the 1997 regulations, compared with an ex-ante expectation of a 20% reduction. The data were normalised for floor area and weather effects, and the study attempted to ensure that the two samples were as alike as possible based on location, house type and number of bedrooms. However, it did not have access to socio-demographic data such as household size or income levels, or to energy use history, and so a self-selection bias may remain. Also, more than 50% of homes used more than one heating fuel and so the study was not comparing changes in total heating energy use. As with Kjaerbye et al, the study does not provide any detail of the assumptions underlying the ex-ante estimate.

It would be useful to be able to compare the results from these bottom-up approaches with a top-down method, to see if there was any degree of consistency. There are no papers in the evidence base looking at Irish or Danish building codes from a top-down perspective. There is however a top-down study from the US: Deason and Hobbs (2012) apply an econometric modelling analysis of State-level energy consumption statistics from 48 continental US States over 12 years to look at the effects of 1992-2006 building codes (which, for modelling purposes, they treat as essentially the same in energy efficiency terms). In contrast to Rogan and O Gallachoir, they find that building codes have reduced overall household energy use by 10%, whereas engineering simulations would suggest 5%. However, the method for this study is not described in sufficient detail for its quality to be assessed, so there is little that can be inferred from this result.

Teidemann (2012) estimates the energy use reductions from the implementation of the 2008 building energy code in British Columbia, using engineering simulations enhanced with data from site audits. Estimated savings in heating energy use of 3-5% are reported but these are not compared with what was expected from the code.

The main potential explanation of differences between ex-ante estimates and achieved savings offered by the literature concerns the extent to which the energy efficiency requirements of the regulations are installed: i.e. the degree of compliance with the regulations. Rogan and O Gallachoir (2011) studied Building Energy Report data for the homes built under the 2002 regulations and found that the certified level of energy efficiency of the dwellings as constructed was 11.7% lower than the requirements of the regulation, suggesting that a significant proportion of the difference between the calculated and measured energy use reduction in this case could be due to compliance issues. Tiedemann reports using a compliance rate of 63%, based on on-site investigations in 187 dwellings. However, it is not clear what 'compliance' represents here: the paper reports that, on average, 63% of dwellings were compliant but does not explain what degree of failure to meet the code classes as 'non-compliant'.

The concept of 'prebound' (Sunikka-Blank and Galvin, 2012) discussed earlier is also relevant here: if householders generally use less energy than an engineering estimate would suggest, but if the gap between the estimate and reality closes as the calculated energy efficiency of a home increases, then the difference between energy used in homes built to successive sets of building regulations will be less than calculated energy efficiencies would suggest.

There is evidence that in some countries modelled ex-ante estimates of savings routinely have correction factors applied to account for compliance levels. For example, in Germany 5% non-compliance is assumed (BMWi, 2011) whilst in the UK, the level is assumed to be 10%[43].

4.2. Energy labelling of buildings

There is very little evidence in the peer-reviewed literature on the effects of building energy labels, with only two papers identified. These suggest little overall effect of certificates in isolation, but that a significant portion of energy saving potential might be accessed if certificates are provided to people who are already interested in saving energy.

Kjaerbye (2009) reports on an evaluation of the effects of home energy labelling in Denmark. Housing energy labels were introduced in Denmark in 1997 and initially were not accompanied by any supporting information or incentive schemes. The requirement for the label was not strongly enforced and hence a significant number of homes sold after the introduction of the label did not actually have a label. The study uses propensity score matching to compare energy use in homes with a label to energy use in homes without; all homes in the sample had been sold within the previous four years. The sample, of just under 4,000 properties, included annual gas use data from utilities and a range of home and household characteristics on which matching was based. The study found that, in most cases, the presence of the energy label had no statistically significant effect on energy use. The author reports this result to be consistent with a number of previous studies using different methods[44]. However, the author also notes that the results are potentially only relevant for gas-heated properties, which are only 25% of the total in Denmark, and also that further investigation would be needed to check the extent to which the matching process used had corrected for differences between the labelled and control groups, because descriptive statistics do show key differences between the groups. A further point to note is that the study looked at a period relatively soon after the introduction of the label: the effectiveness of the programme may well increase over time as market take-up increases, and the relatively slow turn-over in building ownership could mean that the full effect takes many years to be seen.

The only other study of energy label effectiveness found in this review involves a self-selecting group of households in Germany. Herppich (2011) presents an internal evaluation of a utility scheme offering 500 free energy efficiency certificates. The energy efficiency certificates provided to the 500 recipient households identified energy saving potentials of between 20% and 38%. The study surveyed the recipient households to find out which measures they had already implemented, which they were planning to implement, and why. Forty percent of recipient households completed the survey. These responses suggest that approximately 20% of the identified savings had been realised and that a further 40% would be realised if planned investments happened. The scheme received over 10,000 enquiries and it is not clear how the 500 recipients were selected from these 10,000 initially interested households. The author suggests that they are considered 'sufficiently representative' of average households but does not address the issue that they have selected to participate and that the study results come from the 40% who also chose to respond to the survey.

4.3. Appliance market transformation activities

Evaluations of appliance market transformation activities in the time period covered by this study are concentrated in North America and in economies in transition. The latter have been excluded from this analysis as their results are not transferable to a UK context. In addition to three North American studies, an evaluation of incentives for efficient appliance purchase in Japan is also reported here. Earlier studies, reported in ECEEE conference proceedings for 2001, 2003, 2005 and 2007 were reviewed in an attempt to close the gap in information on

43 Pers comm, Hunter Danskin, Head of Technical Energy Analysis, Department of Energy and Climate Change.

44 These previous studies are written in Danish and not reported in peer-reviewed journals or conferences; hence are not included in this review.

appliance policies in the EU, and key results are reported here. Whilst there is little quantitative information of use (largely because this type of market transformation activity is difficult to evaluate using the methods commonly employed for energy efficiency programme evaluation), results from US Federal standards (Meyers et al., 2003) perhaps give an indication of the level of reduction that might be expected from comprehensive standards programmes covering heating, cooling and electrical appliances: the overall effect could be a reduction in household energy use of just under 10% relative to a 'without appliance standards' baseline.

The effects of EnergyStar® labelling on energy use in the US residential and commercial sectors are explored by Webber et al (2000). The paper presents a robust example of a method that combines sales data with engineering estimates of efficiencies. Sales of models of appliances that qualify for the EnergyStar® label are analysed, comparing actual sales with a baseline counterfactual assuming no labelling. Engineering estimates of energy savings per appliance are in some cases enhanced with information about usage patterns, for example through surveys of whether or not energy saving features are enabled by users. However, the authors recognise that the accuracy of the study results could be improved by better knowledge of usage patterns. Results are presented as per unit percentage and GJ annual and lifetime savings: these are useful for evaluating the effects of the programme, but do not provide transferable lessons since they are not discussed in the context of the level of energy savings the programme was aiming to achieve, and overall programme savings are not expressed as a percentage of sector energy use.

The energy savings from US Federal domestic appliance standards are estimated by Meyers et al (2003). Energy use calculations are based on sales data and the average efficiency of the appliances sold. The study concludes that standards taking effect between 1988 and 2007 will reduce residential primary energy demand in 2020 by 8-9% relative to the baseline. The authors identify the projection of energy efficiency levels in the absence of standards as the largest source of uncertainty: this counterfactual is constructed from historical trends and expert judgement on the nature of technical change in the absence of standards. Factors taken into account include government and private research and development, utility and State DSM programmes and consumer information programmes such as energy labelling. The authors note that these various drivers of change were relatively weak over the period in question and hence judge that they are unlikely to have underestimated the rate of exogenous efficiency improvement. This approach accounts for exogenous influences on appliance efficiency, and self-selection bias is not an issue for standards since all householders are in effect 'participants'. Free-ridership is not mentioned as an issue, but it is presumably accounted for in the definition of the baseline rate of efficiency improvement, since this is an estimate of what would have happened in the absence of the programme. The main source of potential bias in the results is the extent to which direct rebound in appliance usage patterns is not accounted for (this is not clear from the paper). If direct rebound is not fully accounted for, the direct net savings may be overestimated.

Baillargeon et al (2012) report an evaluation of a utility DSM programme that aimed to increase the uptake of CFL sales in the Canadian province of Quebec. The programme was not conceived as a market transformation initiative and hence market tracking was not built in to programme design. However, a wider market effect became apparent once the programme was operating and this evaluation used mixed-methods ex-post data collection to build an estimate of the evolution of the market in the absence of the programme. The results are presented as per CFL and overall programme GWh net energy impacts, and so are not easy to compare with results from other studies.

Yoshida et al (2010) estimate the carbon emissions impact of a Japanese programme offering 'eco-points' vouchers for purchases of energy efficient appliances. The authors initially attempted to model the effect of the programme using model sales rankings in a consumer utility function[45], but this did not result in a good fit with sales data. They refined the approach using stated preferences[46] from a consumer survey and were satisfied with the results for refrigerators and room air conditioners. Although the results are presented in terms of overall carbon emissions impact, with little contextual information, there are some potentially interesting findings for this study. In the case of refrigerators, the vast majority of appliances purchased (97%) were replacements and, since refrigerator efficiency has increased dramatically over the past two decades, the net effect of the programme is calculated to be a large reduction in carbon emissions (even when rebound from use of the vouchers is taken into account). However, the result for room air conditioners is very different: only 78% are replacements; the remainder are new additions to the appliance stock. As the energy efficiency of these appliances has only increased slowly over the last decade, the net effect is calculated to be an increase in carbon emissions (although it is not clear how the underlying trend for increased ownership of these appliances has been taken into account).

45 Consumer utility functions are a mathematical expression of how consumers maximise their wellbeing (or utility) by dividing their expenditure between the different classes of goods and services on offer. They reflect consumer preferences for different goods and services, and how these vary with incomes and prices.

46 'Stated preferences methods' involve asking consumers to choose between a series of potential options, and inferring the value placed on each option from these choices (this contrasts with 'revealed preferences', which use actual choices made).

Earlier evidence on the effectiveness of appliance market transformation programmes in the EU includes a review of energy labelling and minimum efficiency standards for refrigeration appliances in the UK and Australia (Lane et al., 2007); a review of the effectiveness of the 1999 EU minimum efficiency standards for refrigeration appliances in Great Britain (Schiellerup, 2001); an evaluation of the effects of labelling in Germany (Schlomann et al., 2001) and an assessment of the effect of EU policies on appliance markets (Bertoldi et al., 2001). These studies examine a range of effects and offer some quantification of the net effect of labelling and/or minimum efficiency standards on appliance energy use. Lane et al examine market trends and combine these with stakeholder interviews to estimate the counterfactual for refrigeration appliance market transformation programmes. They estimate that EU labelling and minimum efficiency standards had reduced UK electricity use by household refrigeration appliances by 2TWh per year by 2006 (i.e. reducing household electricity demand by roughly 2%). This result for one class of electrical appliance is not inconsistent with the 10% estimate for the combined effect of a range of appliance efficiency standards in the US, reported by Meyers et al. Schiellerup, Schlomman et al and Bertoldi et al focus on the sales weighted efficiency of appliances and demonstrate that the introduction of labels and of standards coincided with significant increases in the energy efficiency of appliances sold, but do not attempt to separate the effect of these programmes from other influences on efficiency.

4.4. Investment and refurbishment programmes

Evidence in this area covers both general investment programmes and those targeted at low-income households. The majority of the peer-reviewed evidence comprises reports on utility-run general programmes, together with a mixture of utility and government low-income programmes.

The peer-reviewed literature in the period focused on for this study offers little evidence from utility-run programmes in the US. This illustrates a clear limitation of this study, since evaluation of these programmes in the US has resulted in a significant body of literature in recent years, albeit not peer-reviewed[47]. This gap in the peer-reviewed literature potentially reflects the long timeframe over which utility programmes in the US have been operating: it is possible that peer-reviewed literature from an earlier period would cover these programmes (see, for example, Hirst et al (1985)). However, in addition to being outside the scope of this study, these studies would be difficult to compare with more recent evaluations since factors such as the design of programmes, the underlying market conditions and the starting level of energy efficiency in the housing stock are likely to have changed significantly.

More recent evidence from North America in the peer-reviewed literature concerns alternative, top-down approaches to evaluation, and these are included here.

47 See, for example, http://www.calmac.org/ for an indication of the volume of material from just one US State.

Bottom-up[48] studies of utility- and government-funded investment programmes

As detailed below, the peer-reviewed evidence offers no clear picture of how net direct energy savings in participant households relate to ex-ante engineering estimates for general investment programmes: there is a consensus that they are significantly lower, but estimates of the proportion of theoretically possible savings that is achieved range from 44% to 75%. Taking an alternative, macro-level approach provides a different perspective, but no clearer answers.

As part of an analysis of the effectiveness of the Danish Energy Efficiency Obligation, Bundgaard et al (2013) conducted a case study in the residential sector to determine whether statistically significant net savings could be identified. The study was of a sample of 331 homes in one town, with district heating: 166 of these homes had received energy efficiency-related subsidies from their utility; 165 (the control group) had not. The group receiving subsidies were found to have reduced their energy use in comparison with the control group, but by only 44% of the gross saving level reported by the programme. This is only a small sample; also, although participant and control groups had comparable pre-programme district heating energy use, it is not clear whether self-selection issues are fully dealt with in the methodology, and the potentially confounding issue of non-participant spillover is not addressed. Hence perhaps conclusions from this study should be limited to the idea that actual savings will only be a portion of the gross savings calculated from a simple engineering approach.

Rosenow and Galvin (2013) review evaluations of the UK's Energy Efficiency Commitment schemes[49]. Energy savings are reported only as total programme lifetime TWh savings and hence cannot easily be compared with those from other programmes. The paper does however also note cost-effectiveness: the programmes are estimated to have delivered energy savings at a cost to the energy companies of €0.007/kWh[50]. The authors note that the programme evaluations include adjustments to engineering estimates of energy savings, based on observed savings within a sample of participants. These aim to account for rebound and prebound effects, and the extent to which installed technologies do not perform to the level of energy efficiency that they theoretically should reach (e.g. because of installation errors). The estimates are also adjusted for free-ridership, based on historical rates of measure uptake.

Scheer and Clancy (2011) report on an evaluation of Ireland's Sustainable Energy Authority Home Energy Saving residential retrofit scheme, which provided grants of typically 30% for a range of heating and insulation measures. It is difficult to compare this with utility-run schemes as it is implemented by an Energy Agency and is therefore potentially operating within different constraints and to different aims. However, it offers another view of the extent to which investment programmes deliver theoretically achievable energy savings. The study took a difference-in-difference approach, and used billing analysis to compare participant energy use in gas heated homes with that of a control group matched on the basis of a range of factors considered to affect energy use. The results suggest that the measures have led to a 22.4% reduction in gas use in participant homes relative to the control group. This is 25% lower than engineering estimates would suggest. The study required consent for data use from households that had participated in the scheme and had also paid a small sum for before and after energy ratings; this resulted in a sample of 216 households from the 75,000 that had taken part in the scheme. The sample, although well matched to the control group, was not representative of the population as a whole: it contained a high proportion of retired households; the required 70% householder contribution probably resulted in an under-representation of low-income households; and the requirement for before and after energy ratings may have resulted in a more than averagely energy-aware participant group. The potential effect of non-participant spillover is not considered.

Low-income programmes

There are a small number of studies in the recent peer-reviewed literature[51] looking at low-income programmes, but these are very diverse in their nature and aims, and do not produce an overall picture of likely effects of this type of programme.

The effects on space heating energy use of the UK's main government-funded programme for low- income households, Warm Front, were evaluated by Hong et al (2006). Cross-sectional and longitudinal data on fuel use and internal and external temperatures were collected and the latter used to normalise fuel use, thus excluding the effect of comfort taking (through increased temperatures) on measured fuel use. The study found that loft and cavity wall insulation measures reduced normalised fuel use by between 10 and 17%, whereas model predictions would suggest a 45-49% reduction.

48 'Bottom-up' studies are those which estimate energy savings on the basis of changes in each participant household; in contrast, 'top-down' methods involve estimation from macro-level data such as changes in total household energy use in a given geographical area.

49 These are utility-implemented programmes, responding to a regulatory requirement to meet an energy efficiency target; they followed the Energy Efficiency Standards of Performance and preceded the Energy Company Obligation schemes described in Box 1.1.

50 Householder investment costs are not reported and hence not included in the evaluation.

51 Here again, it is likely that a large number of evaluations of low-income weatherisation programmes in the US are not captured by this study, as they have not resulted in recent peer-reviewed papers.

Installation of gas central heating had no impact on overall normalised fuel use, whereas model predictions suggest a reduction of around 43%. The authors suggest a number of potential reasons for the discrepancy: that the engineering model is too simplistic, in particular in the way it deals with ventilation; that the study algorithm to convert temperature readings into whole house temperatures may be inaccurate; that ventilation patterns may change after installations, and that occupants may well be choosing to continue using inefficient heating appliances in addition to or in place of the newly installed central heating systems.

Mowris and Jones (2012) estimate the effects of a 6,500 household comprehensive energy efficiency programme in the US with the aim of demonstrating the cost-effectiveness of the programme. The results, expressed in total kWh savings and benefit/cost ratios, are not useful for comparative purposes. However, the paper also compares savings calculated from before-after billing comparisons for 58 homes and enhanced engineering simulations (based on inspections in 158 properties, lighting data loggers on more than 1,000 fixtures and participant and non-participant surveys) with ex-ante engineering estimates. They find differences, varying for different measures, ranging between +13 to -10%. With small sample sizes, and no controls, it is difficult to draw any conclusions from these results, although the paper does provide some interesting insights into potential reasons for differences, including higher existing levels of insulation than had been assumed, and longer hours of use of CFLs.

Seifreid et al (2009) present the results of a German government pilot of an assistance package for low-income households intended to reduce electricity use. The paper reports an average reduction of 18% for participant households, but a large range around this. As the paper gives no detail of the estimation method (its focus is on methods of estimating cost-effectiveness for a given level of energy saving) it is not possible to assess the robustness of this result.

Top-down studies of the effects of investment programmes

An alternative approach to estimating utility programme effects, looking from the perspective of overall portfolio or policy effects, has been proposed by a number of authors (Loughran and Kulick, 2004, Rivers and Jaccard, 2011, Dulleck and Kaufmann, 2004, Horowitz, 2007). In general these studies tend to look at the economy as a whole rather than just the household sector, and offer only an overview of the effects rather than any detailed information about relative effectiveness of different programmes or measures. However, it is worth considering whether their results can add anything to the information from single programme evaluations, particularly since their headline results suggest that energy savings from energy efficiency programmes may be significantly lower than individual programme evaluations indicate.

Looking at the effects of US utility DSM programmes, Loughran and Kulick (2004) econometrically model the difference between electricity growth in areas with utilities that have DSM programmes and those without. Using data from 324 utilities across a period of 11 years, they estimate a statistically significant net effect that reduces retail electricity sales by between 0.4 and 1.2%. This compares with an average net reduction of 1.8% estimated by utility evaluations.

The authors suggest that the difference in the estimates is largely due to an underestimation of free-ridership in the utility estimates. They offer suggestions for the drivers of high free-ridership, including the combination of appliance stock turnover and exogenous technological improvement. However, they do not consider specific sectors in detail and hence whether or not the level of free ridership is as high for end uses such as home space conditioning (where the stock turnover, in terms of the building envelope, is very different from that of many electrical appliances). Loughran and Kulick also do not consider the potential confounding effect of non-participant spillover, which may have raised the level of energy efficiency in areas without DSM programmes and hence led to an underestimate of the effect of programmes on energy use.

A more recent paper looking at the situation in Canada (Rivers and Jaccard, 2011) uses significant inter-temporal variations in utility spending on DSM as a quasi-experiment, by modelling how energy use varies with the variation in DSM spending. The authors model the effect of programmes on economy-wide energy use, using a partial adjustment model[52] and find no statistically significant effect. However, this study is based on a relatively small dataset and the authors acknowledge that they had to make quite a number of simplifying assumptions. Since the effect that they are looking for is relatively small compared with total energy use, it is perhaps not surprising therefore that they do not find a statistically significant effect. This study also does not take account of non-participant spillover, although it acknowledges that this may have a confounding effect.

Dulleck and Kaufmann (2004) present a macro-level evaluation of the impact of an Irish Energy Supply Board (ESB) utility DSM programme, involving customer information, small financial incentives for CFL purchase and direct supply of energy efficiency measures. Using a 'traditional' econometric model of electricity demand with the addition of a dummy variable representing the DSM programme, the authors estimate that the long-run effect of the programme was a reduction in overall household

52 Partial adjustment models are used to allow for effects that may occur after a time lag following an intervention

electricity demand of 7%. There is little information about the quantity and quality of data used in the modelling and so it is not possible to judge the robustness of this result. However, the authors argue that the ESB had genuine incentives to ensure that the programme was successful, unlike many utilities running DSM initiatives, and hence the finding of a larger net reduction than in some other macro-level studies is not surprising.

Like Loughran and Kulick, Horowitz (2007) examines the effect of energy efficiency on energy use by comparing the situation in different areas of the US. Using large datasets collected by the US Energy Information Administration, this study characterises US States by the degree of commitment to energy efficiency over a period of time, and models energy use in the residential, commercial and industrial sectors in States with strong or moderate commitment and in those in States with weak commitment.

The initial results of the study suggested that in the residential sector (unlike in the commercial and industrial sectors) strong or moderate commitment to energy efficiency seems to lead to a net (9%) increase in energy demand compared with the counterfactual of what would have happened in these States had there been weak commitment to energy efficiency.

However, the study then goes on to analyse the results in more detail and finds that the nature of householder response to determinants of energy demand in 'weak commitment' States and in 'strong/moderate commitment' States becomes much more similar over time. The author suggests that this is evidence that a non-participant spillover effect is occurring, with energy efficiency programmes in one State having an effect on the market for energy efficiency measures in other States. In the author's view, this effect is potentially large and happens quickly, thus confounding the modelling results based on the 'weak State' counterfactual. This view is based on the model results, but is also supported by further analysis of some key model variables: electricity price elasticity, income elasticity and technical trend variables all change in both strong and weak commitment States in similar ways, which suggests that there is a high degree of spillover between the States. The author points out that the structure of the market for household electrical goods (few major nationwide manufacturers and retailers, and mass media reach) makes this finding plausible.

The paper also includes two further model specifications for the residential sector: one changes the cut-off date between the baseline period and the intervention period, because residential DSM activity in some states began earlier than the originally specified cut-off date and this may have confounded the initial results; the second compares the situation in California (the State with the strongest commitment to energy efficiency) with States with weak commitment. The first of these specifications still finds an increase in energy use in States with strong commitment, but is it much smaller than in the original specification. The second comparison finds a large effect: strong commitment to energy efficiency in California has reduced residential energy use compared with a weak commitment counterfactual by 43%.

To summarise, these alternative approaches offer some interesting results that have the potential to help define a minimum level of effect for programmes, but they are too varied (from a 43% decrease in energy use to a 9% increase) to be particularly useful. The possibility that large non-participant spillover effects exist is an interesting finding. The fact that these studies are potentially confounded by the presence of this large non-participant spillover suggests that the results cannot necessarily be considered more accurate than more traditional approaches. However, they do highlight the fact that the issue of free-ridership deserves more attention, alongside consideration of how to factor in non-participant spillover. This supports recent calls for more focus on studying the market effects of larger scale energy efficiency programmes, an issue that is discussed further in section 4.9.

4.5. Innovative finance

There is only one piece of evidence on the outcome of innovative finance mechanisms that offers sufficient information to allow assessment of robustness. However, this does report on one of the longest-running European schemes: the KfW CO2 Building Rehabilitation Programme (CBRP) in Germany[53].

Rosenow and Galvin (2013) examine the extent to which evaluations of the CBRP take into account the various elements of the evaluation problem. They find that the reported energy savings, from programme activities in 2007, of an average of 54% of pre-refurbishment consumption do not take into account rebound or prebound effects and that evaluations also do not mention the issue of free-ridership. The authors use values for average rebound and prebound reported in an earlier study of German data (Sunikka-Blank and Galvin, 2012) to adjust the savings estimate, resulting in a proposed average saving of 27% of pre-refurbishment energy use. They review the available information on free-ridership and conclude that a minimum of 11% of households taking out loans would have carried out the energy efficiency work without the added CBRP incentives. As mentioned previously, the estimates of prebound and rebound used here are based on a limited amount of evidence but, together with the free-ridership estimate, they do suggest that evaluations may be significantly overestimating the effect of this programme.

53 The CBRP offers long-term, fixed-rate low interest loans through the German State bank KfW. The loans are to support energy efficiency work during general refurbishment projects or construction of new buildings. The loans are supported by subsidies for the achievement of certain levels of energy efficiency and by general promotional work.

4.6. Information and advice

The evidence about information and advice provision covers a range of approaches, but is skewed towards basic information rather than more in-depth advice. There is very little quantification of effects, and what is presented is not necessarily particularly robust. The methods used here tend to be less robust than for other types of policy, with small sample sizes and reliance on surveys with little reflection on the likely accuracy of responses provided.

Diffney et al (2013) evaluate an Irish government advertising campaign to encourage energy efficiency actions, focusing on the elements that encouraged reductions in heating energy use. The campaign included TV and radio adverts and leaflets from energy companies to their customers. As this was a national campaign, use of a control group was not possible. A regression analysis using before and after gas consumption demonstrated a short-run effect of the leaflets from utilities, but this decayed rapidly: householder surveys to some extent support this finding, since responses suggested an increase in awareness of actions but not discernible effect on self-reported behaviours. The authors mention the potential for the use of multiple heating fuels, but describe this as 'limited'; they note that the possible effect of other ongoing programmes is not taken into account, but justify this on the basis that these are relatively small-scale and also that programmes targeting low-income households will be focused on homes without gas heating, as few low-income households use gas as their main heating fuel.

Murray (2010) reports on an in-house evaluation of an Energy Saving Trust-implemented UK national advertising campaign. The effect on energy saving actions was estimated using a door-to-door stratified random survey, designed to be representative of the UK population and carried out three months after the four week campaign of TV, radio and online adverts and PR activity. The survey asked about energy saving actions and householders who recalled the advertising campaign were also asked about the extent to which it influenced these actions. Additional questions about frequency of actions were asked in an attempt to mitigate the social bias that might influence survey answers. Self-reported actions were translated into energy savings using engineering estimates for investments and government defined estimates for behaviour changes. The evaluation finds the programme to be very cost effective, in terms of the consumer bill savings delivered by the advertising. However, there is no discussion of the robustness of householder estimates of programme influence, or of the methods used to translate actions into energy saving estimates.

Evaluations of more tailored advice in particular seem to rely on very small sample sizes. For example, Rowlands and Hawthornethwaite (2013) use multiple surveys and hourly metered electricity use data to examine the effects of energy audits, but their sample is restricted to just 17 self-selecting households in Ontario, Canada. Hence, although the survey returns suggested that households completed on average 48% of all audit recommendations and an average weather corrected electricity use reduction of 17% was recorded, the range around this average was, unsurprisingly, large and little can be inferred from these results. Similarly, Tiedemann (2004), exploring an alternative approach to the provision of more tailored information in an in-house evaluation of a BC Hydro programme providing an online tool, works with billing information and telephone surveys for a sample of 68 participant households and 63 control group households. Whilst participant responses to the surveys give a potentially interesting qualitative insight into what programmes like this can influence (e.g. a relatively high proportion of insulation-related decisions but a relatively low proportion of window-related decisions are attributed to the programme), it is unlikely that the energy savings outcomes are sufficiently large to be robustly estimated from results in this number of households.

4.7. Smart metering and billing feedback

Billing feedback is the programme type that has received the most attention in the peer-reviewed literature in recent years. This type of approach has only relatively recently been implemented on a large scale and this, in combination with the availability of smart meter data, has allowed experimental approaches to the study of its effects. Most reports of large scale experimental trials concern programmes implemented in the US, where smart metering is far more prevalent than in Europe, and all but one concern programmes implemented since 2007.

Allcott (2011) reports on a very large scale study involving detailed statistical analysis of the effects of 17 different programmes run by OPower[54] for a range of utility clients in the US. These programmes are billing feedback via monthly or quarterly reports for electricity use, and include neighbour comparisons and injunctive norms to describe the household's performance. The reports also offer tailored tips for electricity saving. The programmes nearly all had an experimental design that enabled construction of robust control groups for the analysis, and were based on an opt-out design (with low opt-out rates) so should reflect population average effects. Participants and control were balanced in terms of pre-programme energy use. In total around 600,000 households were involved as participants or control groups. The study found average electricity use reductions ranging from 1.4 to 3.3% across all the programmes. It also found that savings seem to increase over the first two years of programme implementation and that there is some evidence[55] for partial persistence of savings if feedback ceases.

54 Opower is a private company offering data-based energy efficiency services to utilities.

55 As Allcott notes, this finding is from one study only, and so should be treated as preliminary.

A number of OPower programmes are also discussed in detail by Agnew et al (Agnew et al., 2011, Agnew et al., 2012, Agnew and Gaffney, 2013). All three papers look at the results from a Puget Sound Energy trial, and one also considers outcomes from similar programmes in Massachusetts and Sacramento. These programmes are also covered in Allcott's paper. However, the papers do provide more detailed information about the specific programmes being evaluated. In particular, Agnew et al (2012) report on an experiment to explore persistence, in which the treatment group (of 35,000 households) was split into two after two years of feedback provision, with one group continuing to receive reports and the other not. For the group that continued to receive feedback, there was no statistically significant change in either gas or electricity use between years two and three. However, electricity savings reduced in the group where feedback was suspended, and the difference in year three savings between the continued feedback and suspended feedback groups (2.6% vs 2.1%) was found to be statistically significant. For gas use, there is no significant difference between savings for the continued feedback and suspended feedback groups. The authors speculate that this could be because gas savings are more likely to come from investment in energy efficiency measures whereas electricity savings are more likely to come from behaviour changes, although they do not have any direct evidence to support this.

Ashby et al (2012) also review US programmes and again there is significant overlap with the papers previously discussed. Although the quantitative results reported therefore add little to the evidence base, it is worth noting a number of elements of the qualitative study findings about variance in outcomes: baseline energy use seems to have an impact, with higher energy use households achieving higher percentage savings as a result of feedback; and higher frequency of reports (monthly rather than quarterly) also increases the level of savings (although Allcott concludes that the marginal saving is not justified by the additional cost of more frequent reports).

Maclaury et al (2012), Parker et al (2010) and Mendyk et al (2010) report on a number of alternative approaches to feedback provision in the US (such as in-home displays and websites including forums for customers to exchange ideas on energy saving). The studies are less robust than the others mentioned here, because they are very small scale and/or use less sophisticated analysis methods to estimate savings. However, their findings are consistent with those reported in the larger, more robust studies. There is little in the peer-reviewed literature that directly compares the effectiveness of different feedback methods: this is a gap that probably deserves further attention as there are pilot results reported elsewhere[56] showing, for example, that combinations of different types of message seem to be more effective than any one element of the combination alone.

There is only one report of a large scale European feedback experiment in the literature: Pyrko (2013) describes a large scale 12 month feedback experiment in Sweden, involving 10,000 customers of one utility. The study estimates an average reduction in electricity use for participants of 0.74%, compared with an average increase of 1.5% for the control group. However, there is little information provided about how the participant and control groups were matched or about distributions around these averages other than to note that the range of savings is large. Large scale experiments carried out in the UK between 2007 and 2010 (as part of the Energy Demand Research Project[57]) are not reported in the peer-reviewed literature. However, four of these are included in a 2012 ACEEE international review of findings from studies judged by the authors to be high quality (Foster and Mazur-Stommen, 2012). In addition to the four EDRP pilots, the report covers results from three US programmes (none of which are described in the papers already mentioned above), and one Irish programme. The authors report an average electricity use reduction across these projects of 3.8%[58]. They note the wide variation in results, both between and within the individual pilots and begin to explore the factors that may affect this. In addition to a number of device design and programme process elements, they also note that there may be an effect linked to 'sensitivity' towards real-time energy consumption feedback and that this 'cyber-sensitivity' does not seem to be linked to the usual observable characteristics that affect energy demand. Note that, although these studies are considered good quality, there may be issues with their methods or data quality that are not reported in this overview. For example, Darby et al (2011) note that the EDRP pilots, which they describe as 'trials, not experiments' involved issues with installation of unfamiliar technologies, experimental designs that were not always as robust as possible and a number of data management issues that emerged as the trials progressed, although their results were broadly consistent with those from other trials.

The literature also includes reports on smaller-scale trials in Europe. Schleich et al (2012) and Schleich et al (2011) report on trials of feedback involving 1,500 households in Austria and 600 households in Germany. Average electricity use reductions of 4.5-5% were found, although the authors were only able to conduct cross-sectional comparisons between participant and control groups due to lack of pre-intervention billing history at the time of the study, and feel that the robustness of the results would be improved by difference-in-difference comparisons once sufficient data are available.

56 See, for example, STROMBACK et al. (2011).

57 https://www.ofgem.gov.uk/gas/retail-market/metering/transition-smart-meters/energy-demand-research-project

58 This average excludes the Irish study, as the savings there were much higher than in other studies. It is not clear whether this was due to the dominance of low-income households in the programme or whether there were issues with the experimental method used, leading to an inaccurate estimate.

The EDRP programmes in the UK examined gas as well as electricity savings. Foster and Mazur-Stommen (2012) report that in-home displays were less successful in encouraging gas use reductions than they were for electricity use, but that the installation of smart meters seemed in itself to result in an average gas use reduction of 3%. Drozdowski and Vandamme (2013) report on a trial in 400 households in France, involving smart gas meters and fortnightly energy use reports, which built on the experience from the UK EDRP programmes. The study suffers from self-selection bias as the participants were volunteers and hence the authors take a conservative approach to estimating savings, basing it solely on self-reported actions taken by the end of the 8 month trial period rather than also including any self-reported planned actions. On this basis, they report that smart gas meters and feedback could lead to at least a 0.9% reduction in gas usage.

Although there is an apparent degree of consistency in all these results, with feedback programmes leading to average energy use reductions of somewhere in the region of 1 to 5%, this is actually a fairly large range in a small effect. Also, all the studies report large variances around the averages and there is some focus in the literature on how future work could explore these variations in more detail to enable appropriate targeting of future programmes. Writing about lessons that can be learned from the Energy Demand Research Projects in the UK, Darby et al (2011) note two points of particular interest in planning further work on this type of programme: comparability across trials is difficult because the design of the EDRP programme did not involve the definition of consistent projects that could be compared (an issue that can only be more pronounced if the comparison is between results from different programmes); and many of the approaches trialled included several interventions at one time, making it difficult to develop an understanding of the impacts of each individual intervention.

The experimental approaches employed in these studies will mean that exogenous influences will generally be well dealt with. The estimates produced will include, but not separately quantify, the effects of participant spillover and rebound on the fuel being monitored; however, many of these evaluations look only at one fuel (usually electricity) and so any participant spillover or rebound that affects the use of other fuels in the home will be missed. Treatment of free-ridership varies, depending on the extent to which the control group seems well matched with the participant group and the same comment applies to self-selection. These studies could be subject to the confounding effect of non-participant spillover, although the likely magnitude of this effect is unknown.

4.8. Community-led energy action

There is very little robust outcome evaluation of community-led[59] energy activities reported in the literature. This may reflect the historical lack of priority given to this type of programme; equally, it could reflect the preference for the implementers of such actions to focus on process rather than outcome evaluations, aiming to improve initiatives that they already consider to be effective or to increase their reach. As community energy activities evolve to include elements of investment and financial return, evaluation of realised energy use reductions may become more important to the programme implementers.

Two of the three papers considered here report on 'open house' events, in Australia (Berry and Sharp, 2013) and the UK (Hamilton and Killip, 2009). Both studies were based on questionnaires administered to event visitors. Both focus on perceived usefulness of the event and on stated intentions to take action. Both report very positive results in terms of the extent to which the events inspire and enable action, and Berry and Sharp also conducted a follow-up study where self-reported actions taken matched well with previously stated intentions for the 12-month period after the event. However, neither study is able to verify householder reported actions with any recorded data on changes in energy use.

Ferreira et al. (2009) report on an in-house evaluation of a small-scale intervention delivered by an environmental organisation in Portugal. The intervention focused on delivery of low-cost, simple energy efficiency measures. One focus was standby consumption: the delivery organisation identified the potential for a 5% reduction in home electricity from the reduction of standby. Householders were given advice on how to reduce standby consumption and provided with switched extension leads. Monitoring of specific appliances in a small sample of participant households indicated that 80% of the identified standby savings were being achieved. This is an interesting result, but it is for a small, self-selecting sample, and the study does not consider the extent to which the energy saving actions persist.

59 The programmes reported here are led by non-profit community organisations. However, the term 'community-led energy' can also include local activities undertaken on behalf of the community by local government, and those in which local government and community organisations work together.

4.9. Wider impacts of energy efficiency programmes

The methods described in Section 2.1 are focused on energy savings within household boundaries, and the vast majority of the studies reported above do not consider the wider impacts of energy efficiency programmes. These wider impacts can be split into two main elements: indirect rebound affecting energy use outside the home; and non-participant spillover. Neither are covered comprehensively in this study: the former because it has been extensively reviewed by a UKERC TPA report previously (Sorrell, 2007); the second because there are as yet few studies that examine the effect. However, it is worth summarising the state of knowledge here to help set the context for the evaluation recommendations that are made in Chapter 5.

Indirect rebound affecting energy use outside the home

The potential for rebound effects (both direct and indirect) to reduce or potentially even reverse the energy saving effects of energy efficiency programmes is well recognised. Sorrell's 2007 review remains the most comprehensive summary of the state of knowledge in this area. In it, the author states that 'both direct and indirect effects appear to vary widely between different technologies, sectors and income groups and in most cases they cannot be quantified with much confidence'. However, studies frequently find that economy-wide rebound effects exceed 50%, and this 'should give cause for concern'. More recently, Druckman et al (2011) have explored rebound from a number of household energy efficiency improvement options in more detail. Using household consumption functions to explore changes in spending patterns resulting from lower energy bills, they note that reducing energy use for heating will have a relatively low indirect rebound because heating expenditure is one of the most energy intensive elements of household spending. They estimate that, assuming that fuel bill savings are transferred to other spending categories and/or savings proportionately to overall average proportions of expenditure in these categories, the indirect rebound in this case could be as low as 7%. The latest IPCC work (IPCC WGIII, 2014) concludes that there is no evidence to suggest that rebound effects for buildings energy efficiency are large, noting that in countries with strong policies for energy efficiency in buildings, energy use is decreasing.

Non-participant spillover

Vine and Thomas (2012) note that the overall impact of programmes on national levels of energy use and emissions is increasing and at the same time greater emphasis is being placed on market transformation and changing the social norms related to energy use. These two developments both increase the importance of non-participant spillover effects: as more people explicitly take part in programmes, it is likely that an increasing number of their friends and neighbours will learn about the steps they have taken, and some will take similar actions; and programmes that have an effect on the market for an energy efficient technology will alter the price and availability of that technology for all households, not just those that take part in the programme.

Consequently, policy and portfolio level evaluation is increasing in importance, whilst at the same time programme level evaluation remains relevant to determine the cost-effectiveness of different approaches and to help improve programme design and delivery (Vine et al., 2012). A number of studies of utility DSM programmes conducted at this level were mentioned in section 4.4. However, the methods used in these studies (econometric analysis of billing data with comparisons between different US States) will mean that the effect of rebound is captured in the estimate, but the effect of non-participant spillover is ignored. As discussed previously, Horowitz (2007) identified this as an issue. Writing more recently, Horowitz (2011) argued that neglecting non-participant spillover and broader market effects could lead to a significant underestimation of programme effects (unlike neglecting rebound, which leads to an overestimation). There are a number of recent US studies that explore non-participant spillover in more detail, looking at the mechanisms through which a broader effect occurs (see for example Baillargeon et al, (2012) and Vine (2013)). These do not however attempt to quantify the overall effect in terms of a percentage of the direct effect on participants.

Net wider market effects

What is not clear from a brief review of key literature on indirect rebound and non-participant spillover is the likely relative sizes of these effects, for different types of energy efficiency programme, or the way in which they interact with one another. There is a need for more work on evaluation at this policy or economy-wide level, although it must be recognised that such evaluations will not provide the same degree of confidence in estimates of individual programme effects as some of the methods discussed above.

4.10. Discussion

The peer-reviewed evidence base on the outcomes of household energy efficiency programmes is relatively small in comparison with the total number of evaluation reports that have been produced. For some types of programme (information and advice, and community energy programmes in particular) there is very little robust outcome evaluation reported. For many others there are significant gaps in the useful quantitative information reported. The need to restrict the evidence base to peer-reviewed literature should offer a useful degree of quality assurance but it is also one of the limitations of the study method used here: although including peer-reviewed conference papers has in this case helped to expand the evidence base to include the work of a greater range of evaluation professionals, there remains a very large body of work that is not represented here, particularly for utility and other refurbishment programmes. The remainder of this discussion section refers to the peer-reviewed evidence base, and some of the gaps identified may well be covered in the work not included here. Chapter 5 includes a discussion of the extent to which further work is merited to explore this wider literature.

What we know quite a lot about

The evidence on minimum efficiency standards for buildings, appliance market transformation activities and investment and refurbishment programmes suggest that these all result in reduced energy use, although the level of savings may be lower than ex-ante estimates suggest. Whilst the level of savings delivered depends on programme design details, the order of magnitude of the savings from these types of programme seems to be around 10%. The evidence also offers a consensus on the average effects of feedback programmes (a 1-5% reduction in household energy use), and on the fact that per household savings exhibit a very large range around the average. It also points to a significant difference between simple engineering estimates of the outcomes of programmes that encourage or require investment in energy efficiency technologies (traditional utility programmes, low income programmes, building regulations) and the actual outcomes achieved, but offers little consensus on how large the difference is (estimates suggest that the energy use reductions achieved could be in the range from 20% to 75% of estimated values) and little reflection on the extent to which commonly used correction factors capture this difference effectively.

What we know very little about

This study explored the different elements of a programme's effect that need to be taken into account during an evaluation (for example, spillover and free-ridership), and has identified instances where evaluations do not include consideration of all the relevant elements. However, what is missing from the literature reviewed here is any consensus on the likely magnitude of many of these effects. There are estimates of rebound, but more work is needed to look more specifically at differences across measures and programme types; very little is known about the extent of non-participant spillover, particularly for programmes other than straightforward investment support; and we do not know the likely size, or even perhaps the direction, of the effect of self-selection bias. A related issue is the extent to which the term 'direct rebound' is misused in the literature. It is often used to describe the entire difference between calculated and measured energy savings, when it explains only a fraction of this. More careful use of terminology would help focus effort on what needs to be done to improve estimates and indeed to increase actual achieved savings.

As mentioned above, there is very little information on the outcomes of information and advice programmes, or on community energy initiatives. There is also very little information as yet about the effects of innovative financing programmes.

A key gap in the literature is the lack of any discussion of the 'reach' of energy efficiency programmes, i.e. the proportion of targeted households that the programme will induce to act. Similarly, reporting of results for investment programmes tends not to include discussion of average per household reductions in energy use, and so the depth of the renovations supported cannot be ascertained.

Analysis of process evaluation literature may help to close this gap: one area where the literature review did find some information about reach and depth was for innovative finance mechanisms, and for this type of programme, papers reported elements of both outcome and process evaluation results. For example, Gillich and Sunikka-Blank (2013) examine a PACE[60] scheme in Maine and suggest that early results show the approach reaching a broader range of people than more traditional finance schemes: the authors report that rebates seem to reach the top 20% of households, in income terms, whilst PACE loans seem to be reaching the top 35%. Also, the World Energy Council published an international review of innovative finance for energy efficiency in buildings (Guertler et al., 2013). The authors were not able to assess the robustness of the estimates, since many of the programme reports provided little information about methods and data. However, they present reported average per household energy savings of 1% in New Zealand's 'WarmUp NZ' programme (targeted at increasing internal temperatures rather than saving energy); just under 40% in Estonia's 'Kredex' programme, and up to 66% in Japan's 'Flat35' loans programme. These latter two results begin to suggest that newer finance

60 'PACE' is Property Assessed Clean Energy. This type of scheme has been tried in a number of US States, and involves low-interest loan funding from local government, secured against the property.

mechanisms may have the potential to stimulate greater levels of per household energy saving, although this conclusion is subject to significant uncertainty resulting from lack of information about estimation methods.

Wider economic impacts of programmes are another area needing further work, as Section 4.9 explained.

Evaluation good practice: is there a 'gold standard' and how far from it are we?

As argued in Chapter 2 of this report, the accuracy of evaluation results will in theory be lowest when a simple engineering estimate approach is used and highest when a Randomised Control Trial has been carried out. However, the practicalities of implementing evaluations in the context of a complex social system such as household energy use mean that defining a 'gold standard' is not as simple as using the theoretically most accurate method. Rather, optimal evaluation practice needs to demonstrate an understanding of the elements of the evaluation problem and develop a pragmatic approach to best reflecting these in the evaluation, given the specific circumstances of programme implementation. The need for an evaluation strategy to reflect a wide range of factors and to choose evaluation methods accordingly was discussed earlier in this report.

In many instances where theoretically less accurate methods are used, there may be very good reasons for this: for example, if the aim of the evaluation is to demonstrate to regulators the cost-effectiveness of utility investment programmes which comprise well-understood measures delivering large savings at low cost, the use of suitably adjusted engineering estimates to inform deemed savings levels may be appropriate and all that is justified.

Having said this, there does seem to be a need for more consideration of the potential to use Randomised Control Trials, or quasi-experimental alternatives to these, more frequently (Vine et al., 2014). In addition, studies in the peer-reviewed literature very rarely report the use of multiple evaluation methods or comparisons between study results and those from similar programmes evaluated using different methods. There is also little discussion of the limitations of the methods that are used. This may indicate room for improvement in evaluation practice, or it may be a shortcoming in the peer-reviewed literature only: evaluators guidance and protocols, for example, (Vreuls, 2005, CPUC, 2006) stress the need for multiple methods, and it may be that the fuller evaluation reports in the grey literature include more information on limitations. However, it may be difficult for evaluators to persuade those commissioning evaluations to invest in more robust approaches since, as mentioned above, there is little information about the size and hence importance of some of the effects that are not taken into account in any given evaluation method.

Linked to this, there is no information in the reviewed literature about the cost of evaluations: in practice, governments and regulators will have guides to the appropriate budgets for the evaluation of different types of programme, but it is not possible to discuss the effect of these on evaluation practice when the costs of evaluations are not reported.

Are engineering estimates fit for purpose?

As mentioned above, there are situations where the use of engineering estimates is clearly the most appropriate strategy: this approach can offer cost-effective evaluation that produces 'good-enough' estimates of programme effects for some purposes. It can also make use of pre-determined default values for savings from common measures[61]. Use of these default values not only minimises the costs of evaluations, by doing so it also makes it easier for a wider range of actors to deliver savings schemes: this may be an important policy goal for some schemes (for example the Italian White Certificate scheme (Di Santo et al., 2011)).

However, it is important to ensure that the estimates are sufficiently accurate, as they can influence which measures play a dominant part in investment programmes. The difference between engineering estimates and actual achieved savings from investments in energy efficiency can result from ignoring rebound, incorrectly modelling existing usage patterns, failure during installation to meet technical specifications, unexpected user response to new technologies, or any combination of these effects.

Authors such as Hong et al (2006) note a range of possible explanations for differences found in their studies, but offer no insight into the likely relative weights of each effect. Work has been done in the past, and continues to be done, to refine default values used in utility programme evaluation, and this will include consideration of the magnitude of some of these key effects. For example, the California Evaluation Framework (TecMarket Works, 2004) includes recommendations for net-to-gross ratios[62] for a range of measure / incentive combinations. Changes in these between iterations of the framework are attributed to receipt of more data from more robust evaluations and also to changes in the market for particular measures. Similarly, in the UK, DECC updates adjustment factors for use in investment programme evaluations on the basis of the latest best available data: the latest iterations draw on a large dataset analysed using a difference in differences approach. At the current time, free ridership levels for many of the major measures installed under these programmes are thought to be negligible, because there is already very high take-up of loft and cavity wall insulation, and householders who have not yet invested

61 Often termed 'deemed savings'.

62 In California, net-to-gross ratios are used to reduce gross estimates of energy savings to account for free-ridership; in other locations this adjustment factor may also incorporate estimates of rebound and spillover.

in them are unlikely to do so in the absence of stimulus from a programme. However, overall average adjustments in the range of 50% are needed to account for free-ridership for some measures, and compliance issues, rebound and so on[63].

Work has also been carried out to define best practice ways of monitoring and verifying savings from specific measures, for example in the US Department of Energy's Uniform Methods Project (Jayaweera and Hossein, 2013).

It is likely that the significant expertise represented by this body of work is leading to engineering estimates that do offer a good estimate of the effects of certain programmes and measures, but these estimates tend not to be exposed to review in the academic literature. More open debate of these estimates may perhaps increase confidence in them as a key evaluation method in certain circumstances.

Limits to billing data and the use of control groups

Simply attempting whenever possible to use evaluation methods that are in theory more accurate may not always be the best answer as there are, in practice, confounding factors that can introduce inaccuracies into estimates using these methods.

As an example, there are limits to the accuracy and usefulness of billing data. Many of the studies of billing feedback concentrate on only one fuel, because feedback programmes often only include one fuel. Therefore the study will only capture the effect of the programme on the use of that fuel. Hence the evaluation will miss any elements of rebound that affect other fuels, some elements of unexpected user response to new technologies (for example, changes in use of secondary heating fuels) and will also miss any heat replacement effect caused by the use of more efficient electrical appliances and lighting.

Robust application of statistical analysis of billing data clearly requires expertise in statistics, but the above issues suggest that economics and building physics are also important or at the very least a cross checking of results with enhanced engineering / building simulation estimates. It does not appear from the literature that this happens.

Moving from engineering estimates and before-after comparisons to theoretically more accurate methods requires the definition and use of a comparison group. A key issue here is the potential confounding effect of non-participant spillover. The issue is being recognised (see, for example, (DECC, 2013)) but, as discussed previously, very little is known about the magnitude of this effect.

Data accuracy

Irrespective of the method used, data accuracy and data cleaning will be crucial to the robustness of the results produced. It is interesting therefore that these issues are addressed in only a very small minority of the papers reviewed here. Studies of utility DSM programmes make no comment on the likely accuracy of programme information supplied by the utility; studies using billing data do not specify whether the information is based on actual or estimated meter readings. It is likely that most studies have had to contend with at least some data quality issues, and it is possible that many have dealt with them robustly, but the lack of reporting on these, and on how they were dealt with, means that no conclusions can be drawn about this element of evaluation practice, or about any need for improvements in the data that are available to evaluators.

Assigning effects to multiple mechanisms

The lack of robust assessment of information and advice programmes may in part reflect the difficulty of evaluating these types of programme. The programme is often implemented together with other initiatives, which it is deemed to support. Separating out the effect of the information or advice from the other mechanisms being employed is an issue that is not addressed well in the literature reviewed here. Evaluation theory considers how to assess the overall effect of a set of influences on a household, it does not address how best to assign proportions of this effect to different mechanisms. Vine (2013), discussing market transformation programmes, suggests that the proportion of a change in the market that is due to a particular programme can be explored by examining a series of alternative hypotheses of how the change in energy efficiency or market share may have come about. This can then lead to an estimated range for the proportion of the total effect that may have been caused by the programme in question. Theory-based programme evaluation is increasingly recognised as an important contributor to understanding effectiveness, and it may be that this approach can be used not only to assign a proportion of an effect to a programme, but also to assign proportions of the programme effect to the different mechanisms it is using.

63 Pers comm, Hunter Danskin, Head of Technical Energy Analysis, Department of Energy and Climate Change.

Changing evaluation aims

As has already been noted, the scale of some programmes is increasing and this leads to different evaluation requirements (in the case in point, a need for better understanding of how best to engage householders). There are a number of changes to policy aims that are leading to changes in evaluation design: for example, carbon emissions reduction aims have led to increased investigation of the net, economy-wide effects of energy efficiency programmes.

The need to meet economy-wide targets also leads to a requirement to understand the 'reach' of programmes – i.e. the proportion of energy use they are likely to affect. There is not a great deal of consideration of this element of programme effectiveness in the evaluation literature although a number of initial indications are given in a small number of the papers reviewed: studies of feedback programmes tend to be designed as 'whole population' experiments and hence the findings of a 1-5% average saving are an indication of the overall level of saving across the whole population; the studies on a single tightening of building regulations or the introduction of a series of appliance standards indicate that these sorts of programme lead to overall savings in the region of 7-10% in housing energy use. Evaluations of utility programmes do not generally enable this type of conclusion, because they are aimed at meeting regulatory requirements and hence tend simply to report overall kWh savings, with no reference to the scale of effect that this represents. However, the required data are available: Loughran and Kulick (2004), for example, refer to findings of energy savings that are equivalent to 1-2% of utility sales. Programme reach can be examined in more detail, looking at the types of household that respond to programmes and the depth of saving that each participant household achieves in comparison with identified technical potential. There is little evidence in the literature on either of these aspects, since neither has in the past been an objective of evaluations, but this may change as both the breadth and depth of programme reach become important in the context of climate-related policy goals.

Programmes are becoming more complex, with most employing multiple mechanisms to encourage householder action, and this is reflected in the discussion above about assigning overall effects between these different mechanisms.

One of the most significant changes in programme design in the UK in recent years has been a move away from energy supplier or government funded investment towards mechanisms that support householder investment in energy efficiency. The introduction of the Green Deal in the UK follows a trend of increased use of 'pay-as-you-save' mechanisms in many parts of the world. If householders are to be encouraged to invest in energy efficiency on the basis that they will recoup a financial return, it becomes more important to understand how the effectiveness of energy efficiency investment varies between households rather than simply the average effect achieved. The literature reviewed here acknowledges the extent of variation in results, but offers only initial ideas on the determinants of the variance.

Another significant change in the UK is increased attention on community energy action. As reported above, there is very little evidence in the peer-reviewed literature on the outcomes of community-based programmes. If the level of ambition for this type of programme is to increase significantly, greater attention to outcome evaluation is urgently needed. Similarly, for behaviour change programmes, the literature to date has tended to focus on billing feedback: this focus is likely to have to expand, as a greater range of behaviour-oriented programmes are implemented.

Reporting practices: aiding comparability and transferability

As noted in section 3.4, few papers explain the context for the programme they are evaluating. Without this, it is not possible to assess the extent to which the reported results are likely to transfer to other contexts. At a very basic level, many reports even fail to explain key elements such as whether or not the fuel being measured is used for heating and, if so, whether there are also secondary heating fuels in use.

Equally, energy savings are frequently reported as either per household or overall kWh savings with no reference to pre-programme levels of energy use. Added to this there is often little discussion of the energy goals and objectives of the programme (target savings, or investment levels). Without an idea of the outcome in terms of percentage energy savings and the level of ambition that has generated this, it is not possible to begin to compare results across different programmes (setting aside the issues that differing definitions of net energy savings cause for such comparisons).

Ongoing work to improve comparability of evaluation findings includes the World Resources Institute initiative on greenhouse gas mitigation programme reporting, mentioned previously, and work at Lawrence Berkeley National Laboratory in the US (Hoffman et al., 2013).

4.11. Summary

The evidence base within the peer-reviewed literature demonstrates a wide range of interesting aspects of the energy saving outcomes of energy efficiency programmes, but the answer to the question posed by this study: 'what is the evidence that energy efficiency programmes targeted at the household sector have delivered real energy savings?' has to be: in this sub-set of the literature, it is generally affirmative, although partial, varying in quality, and inconclusive regarding the precise magnitude of the energy savings delivered. The final section of this report offers some recommendations for future priorities that may start to fill in some of the key gaps in this evidence base.

5. Recommendations

This final section of this report presents a number of recommendations for evaluation research and practice, based on the findings of the literature review carried out in this study.

5.1. Evaluation research

This study has identified a number of key gaps in the peer-reviewed literature. These suggest three areas in particular where further research effort may be warranted.

First, there is a need for greater understanding of the importance of some effects commonly not captured in evaluations. Researchers could usefully contribute by considering the likely magnitude of error introduced if an effect is ignored for a particular type of policy. For example, when is non-participant spillover likely to be a significantly large effect that the robustness of control group comparisons is compromised? What market conditions are likely to lead to large free-ridership levels?

Secondly, both the economy-wide impacts of packages of energy efficiency programmes, and the reach and depth of individual programmes, need to be examined more to ensure confidence in the contribution these will make to meeting carbon emissions reduction aims. The first of these elements is a focus in recent literature, but it appears that too little attention is being paid to the second.

Thirdly, newer types of programme will require further attention: community-led programmes, behaviour programmes other than billing feedback and innovative finance options are three obvious examples here.

Alongside these clear gaps, there a number of areas where this study found little evidence but it is likely that gaps could be closed by further work on existing grey literature. This grey literature is extensive and in some cases not easy to access. More work on specific programme types, to review the body of evidence and then subject the findings of the review to peer scrutiny, would be valuable in improving access to the knowledge contained in this literature and in contributing to the debate on which evaluation methods should be used for any given programme. In particular, there is a very significant body of knowledge about the effects of large scale investment programmes that is currently not easily accessible, either because it is spread across thousands of individual programme evaluation reports or because it is not published.

There is a large amount of programme evaluation literature in languages other than English. Work to review and discuss the grey literature should therefore be carried out by multi-lingual teams: multi-national European projects could offer a useful contribution here, enabling learning from programmes in different EU countries to be compared.

This study found very few outcome evaluations that attempted to explore how overall programme effects could be assigned to different mechanisms within the programme (for example, to subsidies for measures or to the information and advice used alongside these subsidies). This may be an area where further work is needed, but equally a review of existing process evaluation literature may offer more evidence.

5.2. Evaluation practice and priorities

Evaluation strategy was discussed in Section 2.4. Current priorities for evaluation work need to be set taking into account the range of factors described there. Given that there is a significant body of knowledge that has not been accessed by this study, there is a limit to what can be expressed here in terms of evaluation priorities, beyond the economy-wide impacts and effects of newer types of programme already mentioned as areas for further research, above. However, some of the shortcomings in the literature reviewed do suggest some areas where evaluation practice may need to develop.

Firstly, it seems likely that there is a need for more evaluations to use multiple methods to estimate programme outcomes. Evaluation guidelines and protocols do recommend such an approach[64], but the extent to which this guidance is followed is not clear. The increasing complexity of programmes and their effects only increases the importance of this cross-checking of results. As programmes become more complex and attempt to reach larger proportions of the population, variation in measured results between different households is another area where much greater understanding is needed. Innovative combinations of smart metering datasets, national energy efficiency datasets (such as the US Energy Information Administration's form 861 data[65] or the UK Government's National Energy Efficiency Data Framework data[66]), and householder surveys are likely to be needed.

Evaluation teams need to be multidisciplinary if all the potential effects of complex programmes are to be understood and adequately reflected in evaluations. Statistical knowledge is vital for robust data collection and handling, but equally building physics, sociology, psychology and economics will be needed to ensure that all the potential reactions of the building and its occupants are taken into account.

64 Vreuls (2005) for example, notes that 'only in rare cases are the results of any one of these approaches to assessing net programme effects on efficiency technology adoption found to be definitive on their own. Therefore it is best to plan to capture information to support at least two, if not all, of the approaches to baseline development…'

65 http://www.eia.gov/electricity/data/eia861/

66 https://www.gov.uk/government/collections/national-energy-efficiency-data-need-framework

The recommendations on evaluation research, above, noted the extent of knowledge not captured in the peer-reviewed literature: evaluation practitioners should be encouraged, including by funders, to expose their results to peer review as the exchange of learning between different programmes that could be facilitated would be of benefit to both evaluators and those commissioning evaluations.

It is virtually impossible to compare the outcomes of different programmes and hence to learn more about what makes a programme particularly effective (or not) in comparison with other similar initiatives. The issue of standardisation of methodologies and reporting practices has been discussed earlier in this report. One further point should be added here: many outcome evaluations report their results as total energy savings. They do not express the results as a percentage (e.g. of per household energy use, or total sector energy use); nor do they relate the outcome to the level of inputs to the programme (e.g. savings per £ invested). A value for total energy savings needs to be reported, for example for the calculation of total carbon emissions reductions, but for results to be compared across programmes, percentage savings and savings per unit input also need to be reported.

References

AGNEW, K. & 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., NIU, M., TANIMOTO, P., GOLDBERG, M. & WILHELM, B. 2011. MO'Power to the customer: an evaluation of a dual fuel home energy reports program. International Energy Program Evaluation Conference. Boston.

AGNEW, K., ROSENBERG, M., TANNENBAUM, B. & WILHELM, B. 2012. Home energy report forms: power from the people. American Council for an Energy Efficient Economy Summer Study. Asolimar, California: American Council for an Energy Efficient Economy

ALLCOTT, H. 2011. Social norms and energy conservation. Journal of Public Economics, 95, 1082-1095.

ASHBY, K., FORSTER, H., CENICEROS, B., WILHELM, B., FRIEBEL, K., HENSCHEL, R. & SAMIULLAH, S. 2012. Green with Envy: neighbor comparisons and social norms in five home energy report programs. American Council for an Energy Efficient Economy Summer Study.

BAILLARGEON, P., SCHMITT, B., MICHAUD, N. & MEGDAL, L. 2012. Evaluating the market transformation impacts of a DSM program in the Province of Quebec. Energy Efficiency, 5, 97-107.

BERRY, S. & SHARP, A. 2013. The role of open house events to improve energy efficiency – reaching the new or preaching to the converted? European Council for an Energy Efficient Economy Summer Study.

BERTOLDI, P., WAIDE, P. & LEBOT, B. 2001. Assessing the market transformation for domestic appliances resulting from European Union policies. ECEEE Summer Study.

BMWI 2011. Second National Energy Efficiency Action Plan of the Federal Republic of Germany. Methodological Accompanying Document. Federal Ministry of Economics and Technology.

BROC, J.-S., OSSO, D., BAUDRY, P., ADNOT, J., BODINEAU, L. & BOURGES, B. 2011. Consistency of the French white certificates evaluation system with the framework proposed for the European energy services. Energy Efficiency, 4, 371-392.

BUNDGAARD, S. S., TOGEBY, M., DYHR-MIKKELSEN, K., SOMMER, T., KJAERBYE, V. H. & LARSEN, A. E. 2013. Spending to Save: Evaluation of the Energy Efficiency Obligation in Denmark. European Council for an Energy Efficient Economy Summer Study.

CABRERA, D., SEAL, T., BERTHOLET, J.-L., LACHAL, B. & JEANNERET, C. 2012. Evaluation of energy efficiency program in Geneva. Energy Efficiency, 5, 87-96.

CLG 2007. Building a Greener Future: Policy Statement.

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

DARBY, S., ANDERSON, W. & 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.

DAVIS, A. L., KRISHNAMURTI, T., FISCHHOFF, B. & BRUINE DE BRUIN, W. 2013. Setting a standard for electricity pilot studies. Energy Policy, 62, 401-409.

DEASON, J. & 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.

DECC 2014. Community Energy Strategy: People Powering Change. London: Department of Energy and Climate Change.

DI SANTO, D., FORNI, D., VENTURINI, V. & BIELE, E. 2011. The White Certificate scheme: the Italian experience and proposals for improvement. European Council for an Energy Efficient Economy Summer Study.

DIFFNEY, S., LYONS, S. & MALAGUZZI VALERI, L. 2013. Evaluation of the effect of the Power of One campaign on natural gas consumption. Energy Policy, 62, 978-988.

DROZDOWSKI, R. & VANDAMME, M. 2013. Smart gas meters: assessment of customer response to improved information about their energy consumption. European Council for an Energy Efficient Economy Summer Study.

DRUCKMAN, A., CHITNIS, M., SORRELL, S. & JACKSON, T. 2011. Missing carbon reductions? Exploring rebound and backfire effects in UK households. Energy Policy, 39, 3572-3581.

DULLECK, U. & KAUFMANN, S. 2004. Do customer information programs reduce household electricity demand? the Irish program. Energy Policy, 32, 1025-1032.

EP 2002. Directive 2002/91/EC of The European Parliament and of The Council of 16 December 2002 on the energy performance of buildings. L1/65. OJEU.

EP 2009. Directive 2009/125/EC establishing a framework for the setting of ecodesign requirements for energy-related products. In: PARLIAMENT, E. (ed.) L285/10. OJEU.

EP 2010. Directive 2010/31/EU on the energy performance of buildings. In: PARLIAMENT, E. (ed.) L153/13. OJEU.

EP 2012. Directive 2012/27/EU of The European Parliament and of The Council on energy efficiency. L 315/1. OJEU.

FERREIRA, F., ANTUNES, A. R., ALVES, F. & RAMOS, S. 2009. EcoFamilies: evaluating and promoting energy savings. European Council for an Energy Efficient Economy Summer Study.

FOSTER, B. & MAZUR-STOMMEN, S. 2012. Results from Recent Real-time Feedback studies. . American Council for an Energy Efficient Economy.

FRONDEL, M. & SCHMIDT, C. M. 2001. Evaluating Environmental Programs: the perspective of modern evaluation research. IZA discussion paper series.

FRONDEL, M. & SCHMIDT, C. M. 2005. Evaluating environmental programs: the perspective of modern evaluation research. Ecological Economics, 55, 515-526.

GILLICH, A. & SUNIKKA-BLANK, M. 2013. Barriers to domestic energy efficiency – an evaluation of retrofit policies and market transformation strategies. European Council for an Energy Efficient Economy Summer Study.

GILLINGHAM, K., NEWELL, R. & PALMER, K. 2006. Energy Efficiency Policies: A Retrospective Examination. Annual Review of Environment and Resources, 31, 193-237.

GREENSTONE, M. & GAYER, T. 2009. Quasi-experimental and experimental approaches to environmental economics. Jnl Environmental Economics and Management, 57, 21-44.

GUERTLER, P., ROYSTON, S. & WADE, J. 2013. Financing energy efficiency in buildings: an international review of best practice and innovation. World Energy Council.

HAMILTON, J. & KILLIP, G. 2009. Demonstration, inspiration … replication? Assessing the impact and limits of social learning from Eco-Homes Open Days in the UK. European Council for an Energy Efficient Economy Summer Study.

HANNA, D. & MARVIN, K. 2013. Control Group Wars - There's more than one way to win the battle. International Energy Program Evaluation Conference. Chicago: IEPEC.

HARTMAN, R. 1988. Self-Selection Bias in the Evolution of Voluntary Energy Conservation Programs. The Review of Economics and Statistics, 70, 448 - 458.

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.

HIRST, E., WHITE, D., GOELTZ, R. & MCKINSTRY, M. 1985. Actual electricity savings and audit predictions for residential retrofit in the pacific northwest. Energy and Buildings, 8, 83-91.

HM TREASURY 2011. The Magenta Book: guidance for evaluation.

HOFFMAN, I., BILLINGSLEY, M., SCHILLER, S., GOLDMAN, C. & STUART, E. 2013. Energy Efficiency Program Typology and Data Metrics: Enabling Multi-State Analyses Through the Use of Common Terminology. Clean Energy Program Policy Brief. Lawrence Berkeley National Laboratory.

HONG, S. H., ORESZCZYN, T. & RIDLEY, I. 2006. The impact of energy efficient refurbishment on the space heating fuel consumption in English dwellings. Energy and Buildings, 38, 1171-1181.

HOROWITZ, M. J. 2007. Changes in Electricity Demand in the United States from the 1970s to 2003. The Energy Journal, 28, 93-119.

HOROWITZ, M. J. 2011. Measuring the savings from energy efficiency policies: a step beyond program evaluation. Energy Efficiency, 4, 43-56.

IPCC WGIII 2014. Chapter 9: buildings - final draft report. Intergovernmental Panel on Climate Change.

JAYAWEERA, T. & HOSSEIN, H. 2013. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. NREL.

JOHNSON, K.-E. 1983. Qualitative response models and the estimation of energy savings from utility conservation programs. Energy, 8, 775-780.

KJAERBYE, V., LARSEN, A. & TOGEBY, M. 2011. Does changes in regulatory requirements for energy efficiency in buildings result in the expected energy savings? European Council for an Energy Efficient Economy Summer Study.

KJAERBYE, V. H. 2009. Does energy labelling on residential housing cause energy savings? European Council for an Energy Efficient Economy Summer Study.

LANE, K., HARRINGTON, L. & RYAN, P. 2007. Evaluating the impact of energy labelling and MEPS - a retrospective look at the case of refrigerators in the UK and Australia. ECEEE Summer Study.

LOUGHRAN, D. S. & KULICK, J. 2004. Demand-Side Management and Energy Efficiency in the United States. Energy Journal, 25, 19-43.

MACLAURY, K., COLE, P., WEITKAMP, E. & SURLES, W. 2012. Lessons from the field: the contribution of active and social learning to persistent energy savings. American Council for an Energy Efficient Economy Summer Study.

MCCARNEY, R., WARNER, J., ILIFFE, S., VAN HASELEN, R., GRIFFIN, M. & FISHER, P. 2007. The Hawthorne Effect: a randomised, controlled trial. BMC Medical Research Methodology, 7.

MENDYK, A., KIHM, S. & PIGG, S. 2010. A reflection of ourselves…How households interact with in-home feedback devices: results from a treatment/control experiment. American Council for an Energy Efficient Economy Summer Study.

MEYERS, S., MCMAHON, J. E., MCNEIL, M. & LIU, X. 2003. Impacts of US federal energy efficiency standards for residential appliances. Energy, 28, 755-767.

MOWRIS, R. & JONES, E. 2012. Moderate Income Comprehensive Energy Efficiency Program Evaluation. International Energy Program Evaluation Conference. Rome.

MURRAY, M. 2010. Evaluation of the effectiveness and impact of energy efficiency advertising campaigns. International Energy Program Evaluation Conference. Paris.

OSMAN, L. M., AYRES, J. G., GARDEN, C., REGLITZ, K., LYON, J., DOUGLAS, J. G., G, K., MILNE, K. & LUDBROOK, A. 2008. The effect of energy efficiency improvement on health status of COPD patients. report to the eaga Charitable Trust.

PARKER, D., HOAK, D. & CUMMINGS, J. E. 2010. Pilot evaluation of energy savings and persistence from residential energy demand feedback devices in a hot climate. American Council for an Energy Efficient Economy Summer Study.

PETERS, J. S. & MCRAE, M. 2008. Free-ridership measurement is out of sync with program logic...or, we've got the structure built, but what's its foundation? American Council for an Energy Efficient Economy Summer Study. Asilomar, California.

PROVENCHER, B., VITTETOE-GLINSMANN, B., DOUGHERTY, A., RANDAZZO, K., MOFFITT, P. & PRAHL, R. Some insights on matching methods in estimating energy savings for an opt-in, behavioural-based energy efficiency program. International Energy Program Evaluation Conference, 2013 Chicago.

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.

RIVERS, N. & JACCARD, M. 2011. Electric Utility Demand Side Management in Canada. The Energy Journal, 32, 93-116.

ROGAN, F. & 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 & EYRE, N 2013. The Green Deal and the Energy Company Obligation, Proceedings of the ICE - Energy, pp. 127-136.

ROSENOW, J. & GALVIN, R. 2013. Evaluating the evaluations: Evidence from energy efficiency programmes in Germany and the UK. Energy and Buildings, 62, 450-458.

ROWLANDS, I. H. & HAWTHORNTHWAITE, J. 2013. Residential electricity audit impact study: an Ontario (Canada) case-study. European Council for an Energy Efficient Economy Summer Study.

SAUSSAY, A., SAHEB, Y. & 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. & 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.

SCHIELLERUP, P. 2001. An examination of the effectiveness of the EU minimum standard on cold appliances: the British Case. ECEEE summer study.

SCHILLER, S. R. 2007. Model Energy Efficiency Program Impact Evaluation Guide. Natinonal Action Plan for Energy Efficiency.

SCHLEICH, J., KLOBASA, M., GOELZ, S. & GÖTZ, K. 2011. Smart metering in Germany – results of providing feedback information in a field trial. European Council for an Energy Efficient Economy Summer Study.

SCHLEICH, J., KLOBASA, M. & GOLZ, S. 2012. Effects of feedback on residential electricity demand – results from a field trial in Austria. International Energy Program Evaluation Conference. Rome.

SCHLOMANN, B., EICHHAMMER, W., GRUBER, E. & STOCKLE, F. 2001. Labelling of electrical appliances - An evaluation of the Energy Labelling Ordinance in Germany and resulting recommendations for energy efficiency policy. ECEEE Summer Study.

SEIFRIED, D., RICHETER, E. & SCHÜLE, R. 2009. Improving energy efficiency for low-income families. European Council for an Energy Efficient Economy Summer Study.

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.

SRC 2001. A European Ex-post evaluation guidebook for DSM and EE service programmes. report to the European Commission SAVE programme.

STROMBACK, J., DROMACQUE, C. & YASSIN, M. H. 2011. The potential of smart meter enabled programs to increase energy and systems efficiency: a mass pilot comparison. VaasaETT.

SUNIKKA-BLANK, M. & GALVIN, R. 2012. Introducing the prebound effect: the gap between performance and actual energy consumption. Building Research & Information, 40, 260-273.

TECMARKET WORKS 2004. The California Evaluation Framework. Report prepared for the California Public Utilities Commission and the Project Advisory Group.

TIEDEMANN, K. 2004. Online audits and energy using behavior. In: MORGAN, K., BREBBIA, C. A., SANCHEZ, J. & VOISKOUNSKY, A. (eds.) Human Perspectives in the Internet Society: Culture, Psychology and Gender.

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, K. M., KUSHLER, M. & PRAHL, R. 2012. Emerging issues in the evaluation of energy-efficiency programs: the US experience. Energy Efficiency, 5, 5-17.

VINE, E., SULLIVAN, M., LUTZENHISER, L., BLUMSTEIN, C. & MILLER, B. 2014. Experimentation and the evaluation of energy efficiency programs. Energy Efficiency.

VINE, E. & 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.

WADE, J., EYRE, N., HAMILTON, J. & PARAG, Y. 2013. Local energy governance: communities and energy efficiency policy. European Council for an Energy Efficient Economy summer study.

WEBBER, C. A., BROWN, R. E. & KOOMEY, J. 2000. Savings estimates for the Energy Star® voluntary labeling program. Energy Policy, 28, 1137-1149.

YOSHIDA, Y., INAHATA, Y., ENOKIBORI, M. & MATSUHASHI, R. 2010. Estimating CO2 Emission Reduction in Eco-Point Program for Green Home Appliances in Japan. Procedia Environmental Sciences, 2, 605-612.

Appendix A

Expert Group Members

Ute Collier, Committee on Climate Change

Hunter Danskin, Department of Energy and Climate Change

Malcolm Keay, Oxford Institute for Energy Studies

Michelle Shipworth, UCL Energy Institute, University College London

Steve Sorrell, SPRU, University of Sussex

Peer reviewers

Wolfgang Eichhammer, Fraunhofer Institute, Germany

Ed Vine, formerly of Lawrence Berkeley Laboratory, USA

Appendix B

Databases searched

- Cambridge Scientific Abstracts (http://www.csa.com/)
- Elsevier Science Direct (http://www.sciencedirect.com)
- Thompson Reuters Web of Knowledge (http://wok.mimas.ac.uk/)
- Worldcat (http://www.worldcat.org)
- Open Grey (www.opengrey.eu)
- GreenFILE (http://www.ebscohost.com/academic/greenfile)

Search terms used

The search terms included in the database searches are given in Table B1. These terms were generally included in a single Boolean string, the database title, abstract and keyword fields were searched, and the following terms were excluded to reduce the number of irrelevant returns: nutrition*, metabol*, potential, renewable, wind, and solar.

A separate set of searches was carried out where the terms 'DSM' and 'demand side management' replaced the energy use terms, and the following terms were added to the list of exclusions: psych*, alcohol*, disorder*, violen*, depress*, drug* and disabilit*. This separate search was necessary to exclude papers from the psychology literature relating to the use of the medical use of the term DSM without excluding papers from the main search that looked at programme evaluation from within psychology and related disciplines.

Table B1

Energy use	Other	Household	Policy	Impact Evaluation
Energy demand	DSM	Household*	Polic*	Evaluat*
Energy use	Demand side management	Residential	Program*	Assess*
Energy savings		Domestic		Impact*
Energy efficien*		Home		
Energy consumption				
Fuel consumption				
Appliance*				

Conference proceedings included

- American Council for an Energy Efficient Economy (http://aceee.org/proceedings)
- European Council for an Energy Efficient Economy (http://www.eceee.org/library/conference_proceedings/eceee_Summer_Studies)
- International Energy Policies and Programmes Evaluation Conference (http://www.iepec.org/?page_id=26)

Appendix C

Paper review matrix

Paper ref:		Does the paper / report provide sufficient detail for quality to be assessed? If yes, complete matrix; if no, exclude from bias analysis and any quantitative summary	Y/N	Average score:
	Does the evaluation demonstrate an understanding of how the programme is likely to affect energy use, and hence seek to collect and use appropriate data?	Is the scale and nature of the evaluation appropriate for the programme size and stage, and level of existing knowledge about outcomes?	Is the choice of evaluation method appropriate for the available data?	Are the limitations of the evaluation acknowledged and, where possible, adjusted for?
Score				
4	Evaluation questions, data collection and analysis methods clearly linked to consideration of how the programme will act to affect energy use	The scale and nature of the evaluation is fully appropriate for the programme size and stage, and level of existing knowledge about outcomes	The choice of evaluation method is appropriate for the available data	Limitations to the evaluation, and hence degree of confidence in the results, is discussed robustly. Adjustments are proposed where possible
3	Evaluation questions, data collection and analysis methods generally cover programme mechanisms, but there are gaps	The scale and nature of the evaluation is generally appropriate for the programme size and stage, and level of existing knowledge about outcomes, but there are some inconsistencies	Data availability to some extent compromises the use of the chosen evaluation method	Limitations to the evaluation are discussed, but there is limited attempt to draw out implications for confidence in the results or to adjust for bias / errors where possible
2	Evaluation questions, data collection and analysis methods cover some of the programme's mechanisms, but there are important gaps	The scale and nature of the evaluation is in some respects appropriate, but there are important inconsistencies	Data availability indicates that other evaluation methods may have been more appropriate	There is partial recognition of the limits to the evaluation, and no discussion of confidence levels / possible adjustment of results
1	There is little evidence to suggest that the evaluation has been designed taking into account the way the programme is likely to affect energy use	The scale and / or nature of the evaluation is inappropriate for the programme size and stage, and level of existing knowledge	The choice of evaluation method is inappropriate for the available data	Limitations of the evaluation are not acknowledged

UKERC
UK Energy Research Centre

11 Princes Gardens
London SW7 1NA
tel: +44 (0)20 7594 1573
email: ukercpressoffice@ukerc.ac.uk

www.ukerc.ac.uk
Follow us on Twitter @UKERCHQ

This report is printed on paper that is FSC accredited.