Demand-Side Policy: Mechanisms for Success and Failure

Author Details
Author: Dr. Peter Warren
Affiliation: University College London (UCL)
Address: UCL School of Public Policy, 29/31 Tavistock Square, London, WC1H 9QU, UK
Email: peter.warren@ucl.ac.uk

Abstract

Demand-side management (DSM) policy evaluations, including energy efficiency and demand response, primarily focus on ex-ante appraisals or ex-post evaluations of impacts, such as energy savings, carbon savings and implementation costs. However, there is a knowledge gap in understanding the mechanisms behind the success and failure of demand-side policies. The paper presents the results of a four-year project to systematically review the global evaluation evidence base in order to identify the key factors for success and failure for different types of DSM policy. The realist synthesis type of systematic review has had limited application in the energy policy field and the research developed a methodological approach to apply it to energy policy analysis. The paper inductively identifies 22 key success factors and 25 key failure factors for twelve types of demand-side policy from a sample of 102 high-quality documents, which cover 690 ex-post evaluations and 66 countries and sub-national states across six continents.

Key Words

Demand-side management policy; energy efficiency policy; demand response; policy success; policy evaluation; systematic review

1. Introduction

The mechanisms behind the success and failure of demand-side management (DSM) policies, which cover energy efficiency, demand response, and on-site back-up generation and storage, is a much-researched area in the literature. Previous research has primarily focussed on policy impacts, such as energy savings, carbon savings and implementation costs (Warren 2015). Although impacts are important for understanding how a policy has performed, they provide limited detail on how and why the policy performed as it did, which is crucial for the design of future policies. In some fields, such as education, social care and public health, theory-based evaluations and realist evaluations are employed to better understand programme mechanisms (Pawson and Tilley 1997, Pawson 2002b, Wong *et al*. 2013). However, it has had limited application in the energy policy field (Warren 2014).

This paper discusses the results of a four-year research project to systematically review the evidence base for demand-side policies. The research aims to identify the mechanisms behind the success and failure of demand-side policies by answering the following research question:

How and why do DSM policies succeed or fail, and what policies have been successful?

The paper covers three main areas: the key overall success and failure factors for DSM policy, the key success and failure factors by DSM policy type, and successful DSM policies. Section two outlines the theory of realist synthesis and defines policy success and failure, section three details the methodology and analytical approach employed to answer the research question, section four presents the results for the three mains areas highlighted above, and section five provides the paper's conclusions.

2. Theory

2.1 Evidence Reviews

Rigorous policy evaluation should include a critical appraisal of the full policy process from the policy proposal through design and implementation to the post-policy evaluation stage. Policy theory-based approaches emphasise the need for evaluation during the implementation stage of the policy (see Rossi *et al*. 2004, Rogers *et al*. 2000). They aim to identify issues that occurred in the design and implementation stages in an iterative process of design, evaluation and redesign (Harmelink *et al*. 2008). The realist synthesis type of systematic review (as developed by Pawson and Tilley 1997 and further discussed in Pawson 2002) extends the thinking behind policy theory-based approaches to synthesise evidence that focuses primarily on the mechanisms behind how and why interventions work (or do not work).

Systematic reviews involve collating and synthesising all of the work that has been done on a particular intervention, trial or programme to better understand what works and what does not work (Petticrew and Roberts 2006, Warren 2014). Systematic reviews are a type of evidence review that utilises systematic techniques, such as employing a transparent search strategy, inclusion criteria and the assessment of study quality (Warren 2018). However, systematic reviews are resource-intensive and usually beyond the scope of policy evaluations sought by governments. Some governments, such as the UK, are increasingly using other types of evidence review, such as rapid evidence assessments (REA), to collect data on the impacts of policies (see UK Civil Service 2014). Unlike systematic reviews, which often require a team of researchers working over a minimum of one year, a single evaluator can undertake a REA in less than six months. REA is a narrower and condensed version of a systematic review (Warren 2018).

Evidence reviews, such as systematic reviews, are commonly confused with literature reviews. The former are methods that aim to collect data for analysis using systematic techniques (such as detailed search strategies, inclusion criteria and quality assessment), whereas the latter do not use systematic techniques and simply aim to identify and appraise the key literature in order to determine current thinking and to identify research gaps (Warren 2015). Figure one summarises the main types of evidence review and their key characteristics. A discussion of the figure is provided in the original source (Warren 2018).

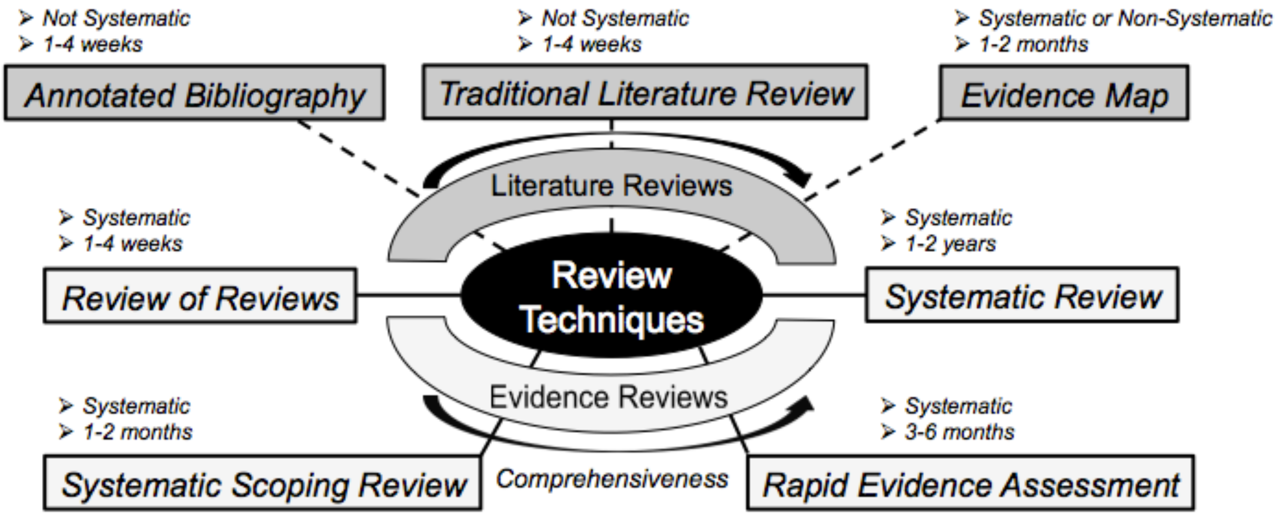


Figure 1: a comparison of the main types of review methods (source: Warren 2018)

Pawson and Tilley (1997) propose the mixed methods quantitative-qualitative realist synthesis type of systematic review in order to focus on mechanisms rather than impacts. Pawson (2002) describes realist synthesis as theory-driven and focused on the underlying programme theory and mechanisms driving an intervention. Comprehensive comparative reviews of the different types of systematic review are provided in Dixon-Woods *et al*. (2005), Snilstveit *et al*. (2012) and Sorrell (2007). Dixon-Woods *et al*. (2005) highlight that the realist synthesis approach tends to treat all forms of evidence as equally authoritative and there is a lack of explicit methodological guidance on how to conduct such analyses in practice. Nevertheless, realist synthesis allows a comprehensive analysis of policy impacts where policies were designed, implemented and evaluated using potentially different methods, by looking at what works and why it works.

Systematic reviews and randomised control trials (RCTs) are two robust methods that are common in other disciplines, such as the medical sciences, but which have had limited application in the energy policy field to date. Although there has been some progress in developing and applying methodological techniques to facilitate a greater use of RCTs for DSM research (such as Gandhi *et al*. 2016, Hahn and Metcalfe 2016, Nicolson *et al*. 2017, Boudet *et al*. 2016, Schultz *et al*. 2015, Nicolson *et al*. 2017), systematic reviews, particularly the realist synthesis type of systematic review, have received much less attention in the field, and there have been growing calls for their application (such as Sorrell 2017, Warren 2014). As such, this paper contributes to filling this research gap by providing new, practical techniques that can be applied in energy and climate policy research. Section 3.1 justifies the adoption of systematic reviews over alternative methods.

2.2 Demand-Side Policy

The research focussed on all aspects of demand-side policy, which covers energy efficiency (achieving the same service with less energy), energy conservation (an overall reduction in energy consumption), demand response (changing patterns of energy consumption in response to price changes or incentive payments), on-site generation (such as on-site renewables and diesel generators), and on-site storage (such as hot water storage tanks and batteries). It is important to note that the latter two categories can be used for demand response purposes (for example, see Siano 2014), but to be classified as demand-side management (DSM), they must be on the demand-side of energy meters. As such, this excludes generation and storage connected at the distribution level. DSM is the umbrella term for all demand-side activities:

"Demand-side management (DSM) refers to technologies, actions and programmes on the demand-side of energy meters, as implemented by governments, utilities, third parties or consumers, to manage or decrease energy consumption through energy efficiency, energy conservation, demand response or on-site generation and storage, in order to reduce total energy system expenditures or to contribute to the achievement of policy objectives, such as emissions reduction, energy security or reducing consumer energy bills." (Adapted from Warren 2018, Warren 2014).

As discussed in the methodology section, the research focused on twelve types of individual DSM policies and nine DSM policy packages. These were determined inductively from the sample of policy evaluations and not pre-defined (as would otherwise be the case in a deductive approach):

- MT: Market transformations (e.g. long-term policies to stimulate the market)
- PBDR: Price-based demand response (e.g. time-of-use pricing)
- IPBDR: Incentive payment-based demand response (e.g. interruptible contracts)
- UO: Utility obligations (e.g. supplier/distributor obligations)
- PS: Performance standards (e.g. for appliances, equipment and buildings)
- LB: Labelling (e.g. for appliances, equipment and buildings)
- IR: Infrastructure rollouts (e.g. smart meter rollouts)
- L&S: Loans and subsidies (e.g. tax incentives and grants)
- UBM: Alternative utility business models (e.g. decoupling policies)
- R&D: Large-scale research and development programmes
- IC: Information campaigns (e.g. marketing campaigns and energy auditing)
- VP: Voluntary programmes

Similarly, the nine DSM policy packages were determined inductively from the sample of policy evaluations and not pre-defined:

- IC/L&S: Information campaigns / Loans and subsidies
- PS/LB: Performance standards / Labelling
- IPBDR/PBDR: Incentive payment-based demand response / Price-based demand response
- UBM/MT: Alternative utility business models / Market transformations
- PS/IC: Performance standards / Information campaigns
- PS/LB/UO/L&S: Performance standards / Labelling / Utility obligations / Loans and subsidies
- VP/L&S: Voluntary programmes / Loans and subsidies
- IC/L&S/MT: Information campaigns / Loans and subsidies / Market transformations

- PS/LB/IC: Performance standards / Labelling / Information campaigns

The focus of the paper is to determine the key mechanisms (also referred to interchangeably as factors) for success and failure for each of these policies. In the context of this research, policy success is defined in relation to the stated success by the evaluators of policy evaluations. There is no single definition for determining the success of a policy (McConnell 2010), and this is evident from the policy evaluation literature. The stated success of a policy refers to the qualitative overall judgement of the evaluator(s) of each evaluation on whether or not the policy has succeeded or failed (McConnell 2010). This includes general statements on whether or not collected or estimated data empirically shows a policy to have been effective, discussions of whether or not the policy met its original overall objectives (McConnell 2010), and statements on how the evaluators of evaluations themselves define policy success. Furthermore, policy success may refer to the degree of unintended consequences, the influence of one policy on another policy (and the degree of policy overlap), the cost-effectiveness of the policy, the degree of actual savings versus modelled savings, and the levels of free ridership and additionality (Warren 2015).

Performance criteria, also referred to as policy impacts, are an important part of determining policy success. Examples in relation to DSM policy include:

- Carbon emissions reduction
- Deferred investment in infrastructure
- Energy bill savings
- Government programme costs
- Overall energy savings
- Peak load reductions
- Utility programme costs
- Consumer active engagement
- Dealing with variable power (such as from wind or solar power)
- Political ease of implementation
- Technology innovation and market development

The definition of ‘policy failure’ refers to policies not performing as well as originally anticipated. The definition mirrors the definition of ‘policy success’, as the primary focus is on the stated failure of policies by the evaluator(s) of each evaluation. Similarly, performance criteria are important in order to ensure that the judgements of evaluators match the data presented. Good examples of papers that use theory-based approaches or elements of theory-based approaches (within a broader policy impacts approach) applied to the evaluation of demand-side policy are Harmelink *et al*. (2008) and Ürge-Vorsatz *et al*. (2007). However, there has been limited research in this area and this paper aims to contribute towards filling this gap.

3. Methodology

3.1 Overall Approach

Due to the exploratory nature of the research question, a pragmatic, inductive approach to science was utilised. This was to ensure that factors for policy success and failure would not be pre-defined and that theory could be built from the evidence. Furthermore, a method was required that could assess the broader, global context on DSM policy, and could aggregate across different country contexts. An alternative approach that this research considered was to undertake interviews with DSM policy experts. However, this would result in data that would be limited by the geographical expertise and location of the experts. Furthermore, the data would be based on expert opinion rather than robust policy evaluations. Systematic reviews are a method that can address these issues.

As justified previously, the realist synthesis type of systematic review was employed to comprehensively capture the global evidence base for demand-side policy evaluation. The details of how the systematic review was specifically employed are provided elsewhere in Warren (2014, 2015); instead, this paper focuses on the analytical techniques and the research results. Nevertheless, in summary, a literature review was first undertaken to identify the main data sources that have published evaluations of DSM policy. This process identified 33 databases and websites, as shown in table one, which were then included in the systematic review.

Data Source Type	Data Source
Academic	Energy Efficiency
	The Electricity Journal
	Energy
	Energy Policy
	Energy Economics
	Energy and Buildings
	Resource and Energy Economics (REE)
	The Energy Journal
	Electric Power Systems Research (EPSR)
Industry	American Council for an Energy-Efficient Economy (ACEEE)
	European Council for an Energy-Efficient Economy (ECEEE)
	International Energy Program Evaluation Conference (IEPEC)
	International Energy Agency (IEA) DSM Programme
	Institute of Electrical and Electronics Engineers (IEEE *Xplore* digital library)
	International Partnership for Energy Efficiency Cooperation (IPEEC)
	British Institute for Energy Economics (BIEE)
	International Association for Energy Economics (IAEE)
	Open Grey
	National Grid
	Association for the Conservation of Energy (ACE)
Government	US Department of Energy (DoE)
	US Energy Information Administration (EIA)
	(Former) UK Department of Energy and Climate Change (DECC) (now the Department for Business, Energy & Industrial Strategy (BEIS))
	UK Office of Gas and Electricity Markets (Ofgem)

	Chinese National Development and Reform Commission (NRDC)
	(Former) Australian Department of Industry (now the Australian Department of Industry and Science)
	Australian Energy Regulator (AER)
	California Public Utilities Commission (CPUC)
	European Commission Department of Energy
	US Federal Energy Regulatory Commission (FERC)
	US Department of Energy (DoE)'s Energy Citations Database
	UK National Audit Office (NAO)
	UK Public Accounts Committee (PAC)

Table 1: the data sources for the systematic review

Returned hits from the searches in the 33 data sources listed in table one were included in the initial sample only if they met the following criteria: they focused on DSM policy mechanisms and not just policy impacts or non-policy aspects of DSM (such as utility-stimulated DSM programmes that were independent of government policies), they were written in English, they were freely downloadable, and they were concerned with government policies and programmes rather than DSM trials, pilots, small-scale research and development programmes, modelling studies of the future potential of DSM, or theoretical aspects of DSM policy. The initial sample was then subjected to a quality assessment scale to determine whether or not the documents were of a high enough quality to be included in the final systematic review sample. The scale is discussed in detail in Warren (2014, 2015) and is not repeated here.

The final sample of 102 high-quality documents covered 690 ex-post DSM policy evaluations. The documents are listed in a separate reference list to the paper's main bibliography. The difference between 'documents' and 'evaluations' is due to some documents evaluating more than one policy or more than one country / (sub-national) state. As discussed in section 2.2, the process was inductive with only the data sources pre-defined. This was to prevent a situation where specific types of previously unidentified DSM policies were excluded from the analysis, despite having a high quality evidence base. The final sample covered 30 countries and 36 sub-national states (including regions and provinces) across six continents. Some sub-national state governments (particularly in the USA) have implemented and evaluated DSM policies independent of national government policy, thus providing an important part of the global policy evaluation evidence base. The majority of the high-quality evaluations in the sample analysed DSM policies implemented in North America, Europe and east-Asia.

As highlighted in section 2.1, there is a lack of explicit methodological guidance on how to conduct such analyses in practice (Dixon-Woods *et al*. 2005), particularly when applied to a field such as energy policy, where evaluations are often conducted using different methods and cover different contexts. The research proposes techniques for data collection and analysis to contribute to filling this methodological gap, which are discussed in this section.

As the documents are of a high quality (having passed the study quality assessment scale mentioned previously) and the evaluators of each evaluation are well qualified to undertake their evaluations, their judgements on overall policy success are considered to be an acceptable indicator, which has been a means of identifying policy success in the evaluation literature (McConnell 2010). Nevertheless, there will always be some degree of subjectivity when using experts' judgements. The new, developed analytical process for the research is visualised in figure two.

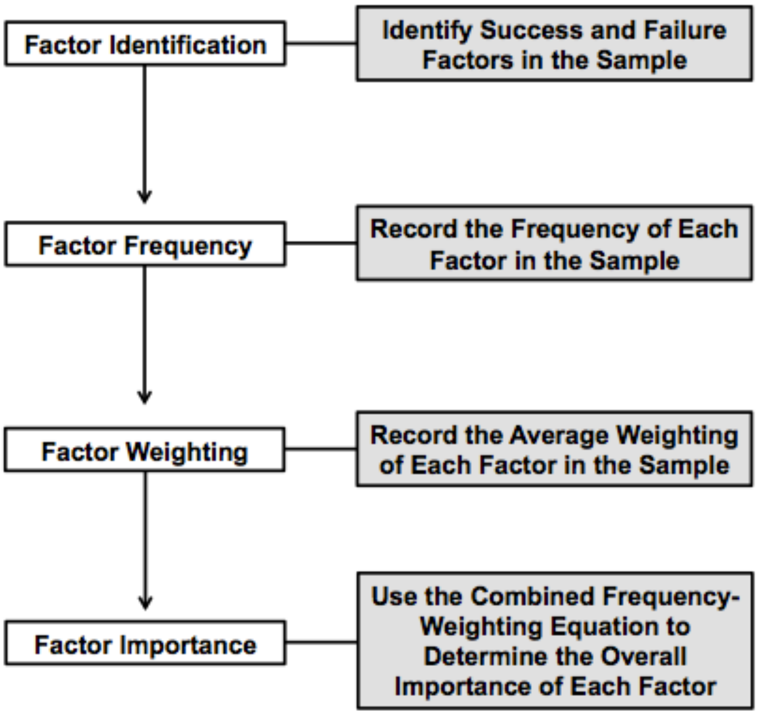

Factor Identification
Identify Success and Failure Factors in the Sample
Factor Frequency
Record the Frequency of Each Factor in the Sample
Factor Weighting
Record the Average Weighting of Each Factor in the Sample
Factor Importance
Use the Combined Frequency-Weighting Equation to Determine the Overall Importance of Each Factor

Figure 2: analytical process for determining demand-side policy mechanisms

The exploration of DSM policy mechanisms is based on identifying the success and failure factors stated by the evaluator(s) of each evaluation in the sample, and then recording the number of times a given factor is mentioned in the sample (frequency) and the stated importance of the factors by the evaluators of each evaluation (weighting). Once the frequency and weighting of each identified factor have been quantified (as discussed in the following sections), the two analyses are combined to give the overall importance of the factor. The combined analysis allows the differentiation of factors that are both frequent and highly weighted from those that are frequent but have a low weighting, are highly weighted but have a low frequency, or are infrequent and have a low weighting.

3.2 Frequency and Weighting Analyses

Factors (also referred to as mechanisms) were determined inductively and are presented in the next section. Success and failure factors were separated, as the results showed that a policy did not necessarily underperform due to the absence of specific success factors, or succeed due to the absence of particular failure factors. As section four highlights, only ten of the factors were the inverse of each other (e.g. 'Political support' (success factor) and ‘A lack of political commitment’ (failure factor)). All of the other success and failure factors were different to each other.

The key strength of frequency analysis is that it shows how widespread the finding is in the sample, and whether or not factors identified for one policy implemented in a particular evaluation and context are present for the same policy in other evaluations and contexts. The frequency threshold shown below was developed to differentiate factors that had a high frequency or a low frequency in the sample.

Factor Frequency Threshold:

1) **High Frequency:** ≥5 evaluations
2) **Low Frequency:** <5 evaluations

The level of ‘5’ evaluations as the threshold was determined inductively by examining the overall average frequency of discussion of each factor in the sample, which also required some degree of judgement (McConnell 2010). As a result, the application of the frequency threshold will vary by sample and study. The main limitation of relying solely on frequency analysis is that it does not identify how important the factors are for a given policy in a given context. Thus, weighting analysis was undertaken to overcome this issue.

In order to calculate the weightings (importance) of factors, a 1.0-3.0 weighting scale was used for each evaluation within each document. The scale is based on the qualitative emphasis that the evaluator(s) of each evaluation give to various factors through the use of specific words, as shown below.

Factor Weighting Scale:

1) ***Score weighting 2.5-3.0 (Crucial):*** the following words are used in direct relation to the factor to strongly emphasis its importance: ‘critical’, ‘crucial’, ‘very important’, ‘necessary’, ‘primary reason(s)’, ‘key’, ‘vital’, ‘central’, ‘essential’, ‘fundamental’, ‘decisive’, ‘significant’ or equivalent
2) ***Score weighting 1.5-2.4 (Some Importance):*** the factor is included at the start of a list of factors and is frequently discussed though it is not strongly emphasised using any of the words for score weighting 2.5-3.0, or it is referred to using phrases such as: ‘quite important’, ‘had some influence’, ‘played a role’ or equivalent
3) ***Score weighting 1.0-1.4 (Small impact but not unimportant):*** the factor is included towards the middle or end of a list of factors without emphasis or discussion or it is indirectly inferred as a factor
4) ***No weighting (Unimportant):*** no weighting is given to the factor

Weightings of 3 (high), 2 (medium), 1 (low) or 0 (unimportant) are assigned to each factor in each evaluation. However, when averages are calculated across policies and countries/states for each factor, figures to one decimal place are used for more detailed comparisons. One limitation of the technique is that the evaluators of evaluations may use language in different ways – for example, one evaluator’s use of the word ‘key’ may be stronger or weaker than another evaluator’s use of the same word. This is a challenge, but the literature is limited in this area, and the proposed technique contributes to filling this methodological gap. Further research should aim to develop this area.

The main limitation of just using weighting analysis is that it does not indicate how widespread the findings are in the sample. Instead, it identifies how important various success and failure factors are in specific contexts. Thus, the weakness of weighting analysis is overcome by undertaking frequency analysis and vice versa, and as such there is strong justification for combining the two analytical techniques in order to identify factors that are both frequent and highly weighted. The following two-part equation was developed to combine the two analytical techniques.

Combined Frequency-Weighting Equation:

1) Frequency-Weighting combined analysis (FW_{pf}) = Policy Success weighting (PS_p) x (Policy Success Factor Frequency (PSF_{pf}) x Policy Success Factor Weighting (PSW_{pf})) / 10

2) Frequency-Weighting combined analysis percentage ($FW_{pf\%}$) = (Frequency-Weighting combined analysis (FW_{pf}) / Theoretical Maximum combined analysis (FW_{pfmax})) x 100%

In notation form:

1) $FW_{pf} = PS_p \times (PSF_{pf} \times PSW_{pf}) / 10$
2) $FW_{pf\%} = (FW_{pf} / FW_{pfmax}) \times 100\%$

Where $_{pf}$ is factor *f* for policy *p*.

The terms in the equation are explained below:

Frequency-Weighting combined analysis (FW_{pf}):

FW_{pf} represents the values from combining the frequency and weighting analyses for a given success or failure factor for a given DSM policy.

Policy Success weighting (PS_p):

PS_p represents the stated success of a given policy through the qualitative judgements of the evaluator(s) of each evaluation as to the overall performance of the policy. To calculate PS_p for each policy, a scale of 1-5 is used (scales of 1-5 are widely used in the field, particularly in surveys, such as the commonly used five-part *Likert scale*, which was developed by Likert, 1932), as shown below.

Policy Success Weighting Scale:

1 = A failed policy that met none of its original objectives
2 = A poorly performing policy that met few of its original objectives
3 = An average performing policy that met most of its original objectives
4 = A policy that performed well and met all of its original objectives
5 = A highly successful policy that performed beyond its original objectives

An average is then taken across the sample for each policy.

Policy Success Factor Frequency (PSF_{pf}):
PSF_{pf} represents the frequency of a given success or failure factor *f* for a given policy *p* in the final sample of 690 evaluations, as determined in the frequency analysis.

Policy Success Factor Weighting (PSW_{pf}):
PSW_{pf} represents the importance of a given success or failure factor *f* for a given policy *p* in the final sample of 690 evaluations, as determined in the weighting analysis.

Frequency-Weighting combined analysis percentage ($FW_{pf\%}$):
Like FW_{pf}, $FW_{pf\%}$ represents the values from combining the frequency and weighting analyses for a given success or failure factor for a given DSM policy. However, it compares the result to the theoretically maximum result that could be achieved (see the explanation for FW_{pfmax} below) and writes the result as a percentage of this. The percentage is used as the final result for determining whether or not a given factor *f* is both frequent and highly weighted for a given policy *p*.

Theoretical Maximum combined analysis (FW_{pfmax}):
FW_{pfmax} represents the theoretically highest score that could be achieved for a given factor *f* for a given policy *p*. This is calculated by multiplying the frequency of discussion of a given policy *p* in the sample with the theoretically maximum possible success weighting of the policy (i.e. 5.0 as per the policy success weighting scale), and then multiplying the resulting value with the overall success weighting of the policy in the sample. The explanation is visualised below.

Theoretical Maximum combined analysis Score (FW_{pfmax}) = Policy Success weighting (PS_p) x (Policy Frequency (P_p) x Theoretical Maximum Policy Success Weighting (PS_{pmax})) / 10

In notation form:
$FW_{pfmax} = PS_p \times (P_p \times PS_{pmax}) / 10$

Where P_p is the frequency of the policy in the final sample of 690 evaluations and PS_{pmax} is the theoretical maximum policy success weighting of 5.0.

In part one of the combined frequency-weighting equation and the theoretical maximum combined analysis equation, dividing the resulting values by ten is undertaken in order to produce a more comparable and manageable scale for categorising success or failure factors. In part two of the combined frequency-weighting equation, the final value is multiplied by 100% in order to obtain a percentage of the theoretically maximum score that is achieved by a given success or failure factor. The level of '5.0%' as the threshold for the combined analysis draws parallels to the frequency analysis threshold and was similarly determined inductively by examining the average combined analysis scores of the various success and failure factors, which also required some degree of judgement (McConnell, 2010). As a result, the application of the threshold will vary by sample and study. The scale shown below was developed to differentiate factors that are both frequent and highly weighted from those that are frequent but have a low weighting, are highly weighted but have a low frequency, or are infrequent and have a low weighting.

Factor Frequency-Weighting Combined Scale:

1) ≥10.0% of the theoretical maximum = Crucial factor
2) 5.0-9.9% of the theoretical maximum = Important factor
3) <5.0% of the theoretical maximum = Unimportant factor

A second level of importance was created in order to identify those factors that are 'crucial', in addition to those that are 'important'. This also better aligns the scale with the three-part (1-3) factor weighting scale shown previously. If the factor achieves ≥10.0% of the theoretical maximum it is considered a 'crucial' factor, if the factor achieves 5.0-9.9% of the theoretical maximum it is considered an 'important' factor, and if the factor achieves <5.0% of the theoretical maximum it is considered an 'unimportant' factor.

The methodological approach and the results were validated in two main ways. Firstly, Multi-Criteria Decision-Making (MCDM) analysis was undertaken with 17 DSM policy experts (from academia, industry and governments) to inductively identify, and then rank, DSM policy success and failure factors. Although the experts were primarily based in the UK and the USA, they are world leaders in the theory and practice of DSM implementation and have extensive international knowledge on the topic. The MCDM analysis was undertaken for two reasons: to cross-validate the results of the systematic review and to perform other analysis to inform the wider research project, which also covered DSM policy implementation and transferability, which are not discussed in the paper (see Warren 2018 (for the results on implementation) and Warren 2017 (for the results on transferability)). The MCDM analysis focused on the overall factors for DSM policy success and failure, rather than attempting to identify the factors broken down by twelve different types of DSM policy from expert opinion. This is due to the limitations of using expert judgement to answer the research question, as discussed in section 3.1, such as geographical constraints. Despite this, it still presents one of the few alternatives to evidence reviews for answering the research question, so it is an appropriate choice for validating the research.

Secondly, the methodological techniques and the results were subjected to review by two academic panels, which were made up of academic experts on DSM and policy analysis. The first panel consisted of five academic experts focusing on the methodological techniques, and the second panel consisted of two academic experts reviewing the research as a whole (these experts are acknowledged in the acknowledgements section). The experts on both panels were selected independently by University College London (UCL), rather than the researcher, in order to eliminate any potential bias.

The next section discusses the results from the combined frequency-weighting analysis in order to firstly identify the overall key success and failure factors across DSM policies and countries/states, and secondly to identify the key success and failure factors for each of the twelve types of DSM policy included in the research.

4. Results and Discussion

4.1 Overall Success and Failure Factors

This section provides a summary of the results, as the results are detailed and cover the success and failure factors for all twelve types of demand-side policy and all 66 countries/states included in the research. Thus, the discussion is split into the overall key success and failure factors when averages are taken across DSM policies and countries/states, and the factors when broken down by DSM policy type. In the case of the latter, the broad groupings shown in section 2.2 are used rather than more specific levels of DSM policy (e.g. price-based demand response rather than critical peak pricing or time-of use pricing, or performance standards rather than building codes or energy efficiency standards for equipment) in order to reduce the boundaries of the research to a more feasible, but still useful, level that can comprehensively cover the global evidence base for demand-side policy. The nature of the research question requires the analysis to span across different countries and continents as well as different types of DSM policy. A larger research project with greater resources should build upon these results by applying the same techniques at a more specific level of DSM policy. For example, Rosenow *et al*. (2016) looked specifically at the implementation of energy efficiency obligations under the EU Energy Efficiency Directive's Article 7 across all 28 EU Member States. Nevertheless, the research focused primarily on conducting an ex-ante appraisal of the patterns of implementation and impacts, rather than undertaking an ex-post evaluation of the mechanisms behind policy success and failure, as is the focus of this research.

22 success factors and 25 failure factors were identified in the systematic review sample, as summarised in figures three and four, which show the frequency (top graph) and weighting (bottom graph) of each factor. As stated in the previous section, only ten of the factors were the inverse of each other.

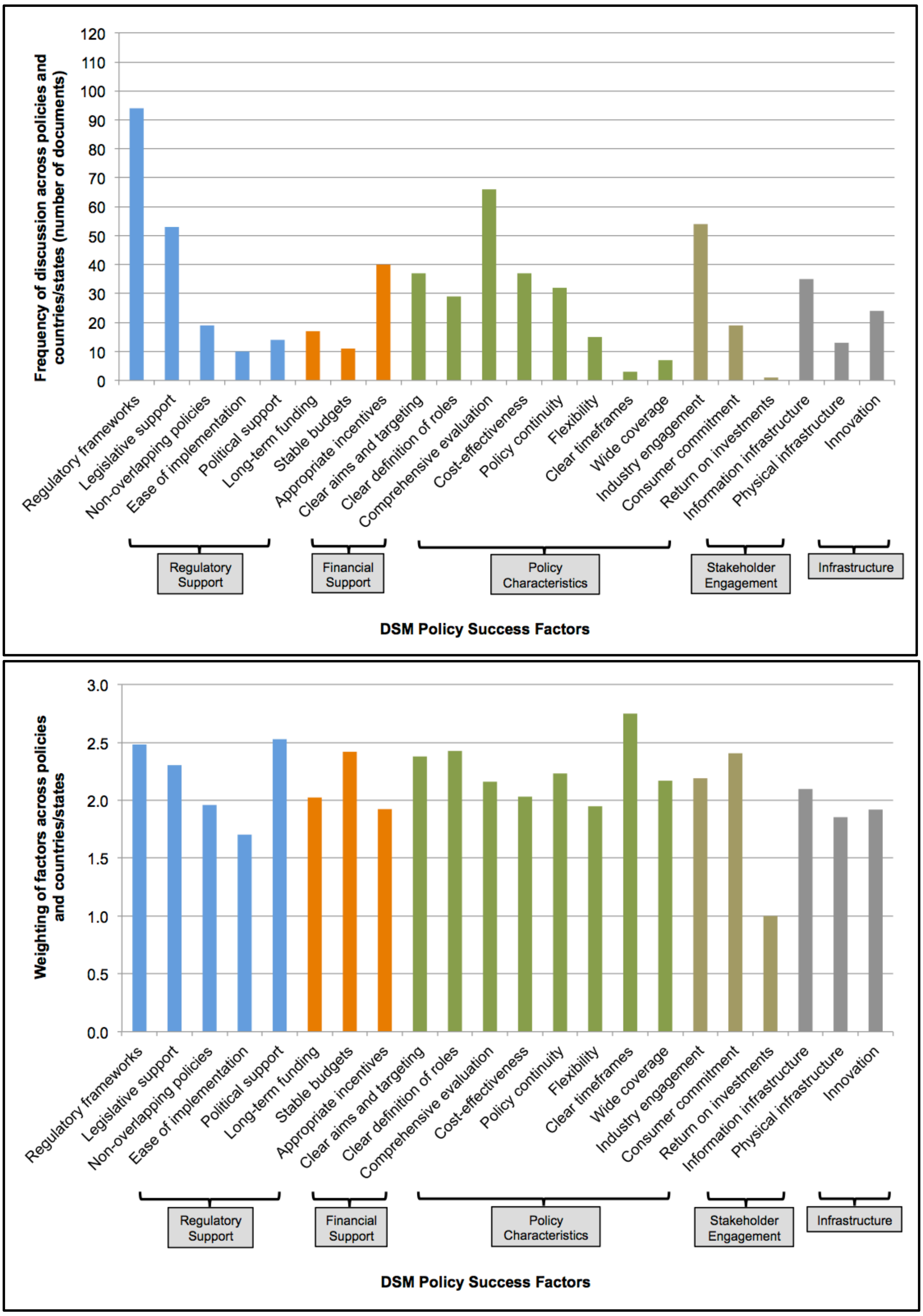


Figure 3: the overall frequency and weighting of DSM policy success factors

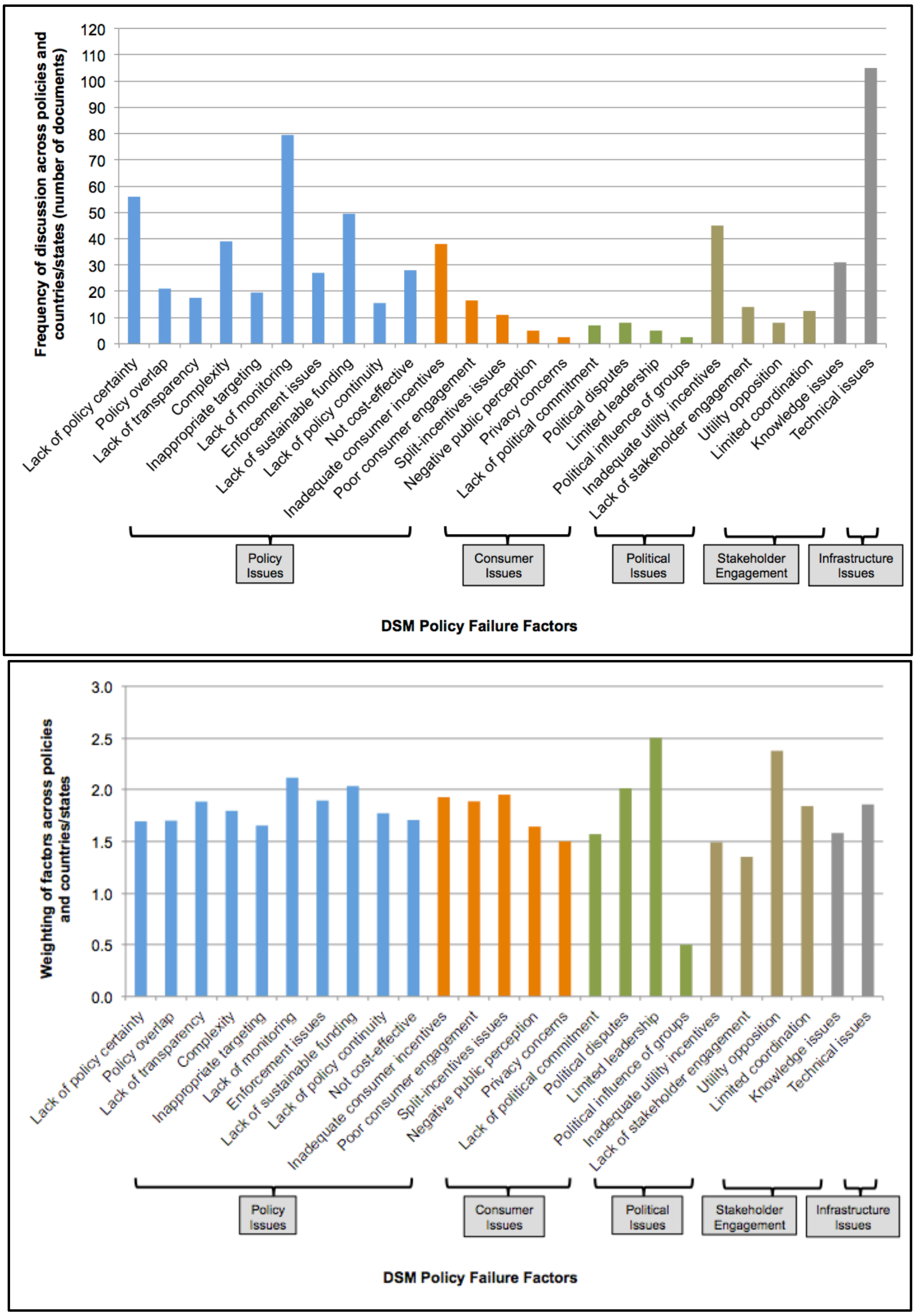


Figure 4: the overall frequency and weighting of DSM policy failure factors

When the individual frequency and weighting analyses are combined using the thresholds, scales and equations outlined in the previous section, the following success factors and failure factors are the most important overall across DSM policies and countries/states:

Overall success factors:

- Regulatory frameworks
- Appropriate incentives

Overall failure factors:

- A lack of monitoring
- Technical issues

'Regulatory frameworks' refers to regulatory rules, government orders, policy frameworks and policy guidance. 'Appropriate incentives' refers to well-designed incentives, which are appropriate to the targeted party. They may be financial incentives or other incentives. 'A lack of monitoring' refers to the lack of adequate resources dedicated to policy evaluation and monitoring during the implementation and post-policy stages. 'Technical issues' primarily refers to programme management and administration issues for relevant parties and government, but it also refers to technological performance problems and a lack of required physical infrastructure (where relevant) caused by programme management issues. The four factors are the most frequent and highly weighted factors in the sample.

4.2 Success and Failure Factors by Policy

The results are presented by individual DSM policy type. Due to limited data, DSM policy packages were not included in this part of the research. Here, the factors for success and failure for the twelve DSM policy types are averaged across countries/states in order to provide a comparative picture on the global scale.

When the combined frequency-weighting analysis equation is applied to the data, the following factors are considered 'crucial' (≥10.0% of the theoretical maximum) or 'important' (5.0-9.9% of the theoretical maximum) factors for the number of different DSM policy types included in brackets:

Success factors by DSM policy:

- Regulatory frameworks (6/12 policies)
- Legislative support (4/12 policies)
- Appropriate incentives (3/12 policies)
- Information infrastructure (3/12 policies)
- Consumer commitment (2/12 policies)

Failure factors by DSM policy:

- Technical issues (6/12 policies)
- A lack of policy certainty (3/12 policies)
- A lack of monitoring (2/12 policies)
- Inadequate utility incentives (2/12 policies)
- Inadequate consumer incentives (2/12 policies)

Figures five (success factors) and six (failure factors) visualise the results. The policy acronyms given in section 2.2 are used. Dark grey boxes with stars represent 'crucial' factors and light grey boxes represent 'important' factors. Here, the importance of regulatory frameworks as the dominant success factor and technical issues (primarily programme administration issues) as the dominant failure factor for half of the DSM policies under examination is clear.

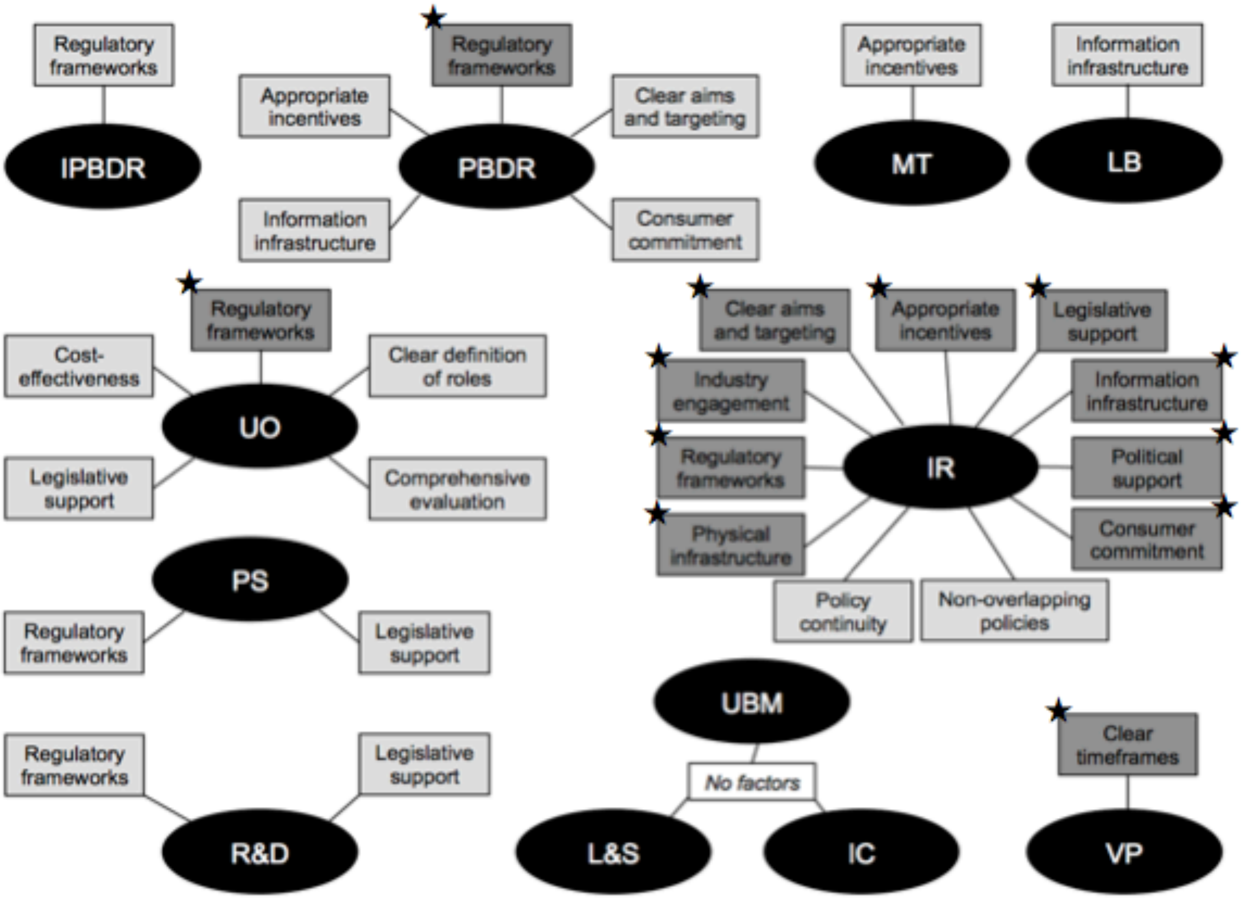


Figure 5: DSM policy mechanisms for success

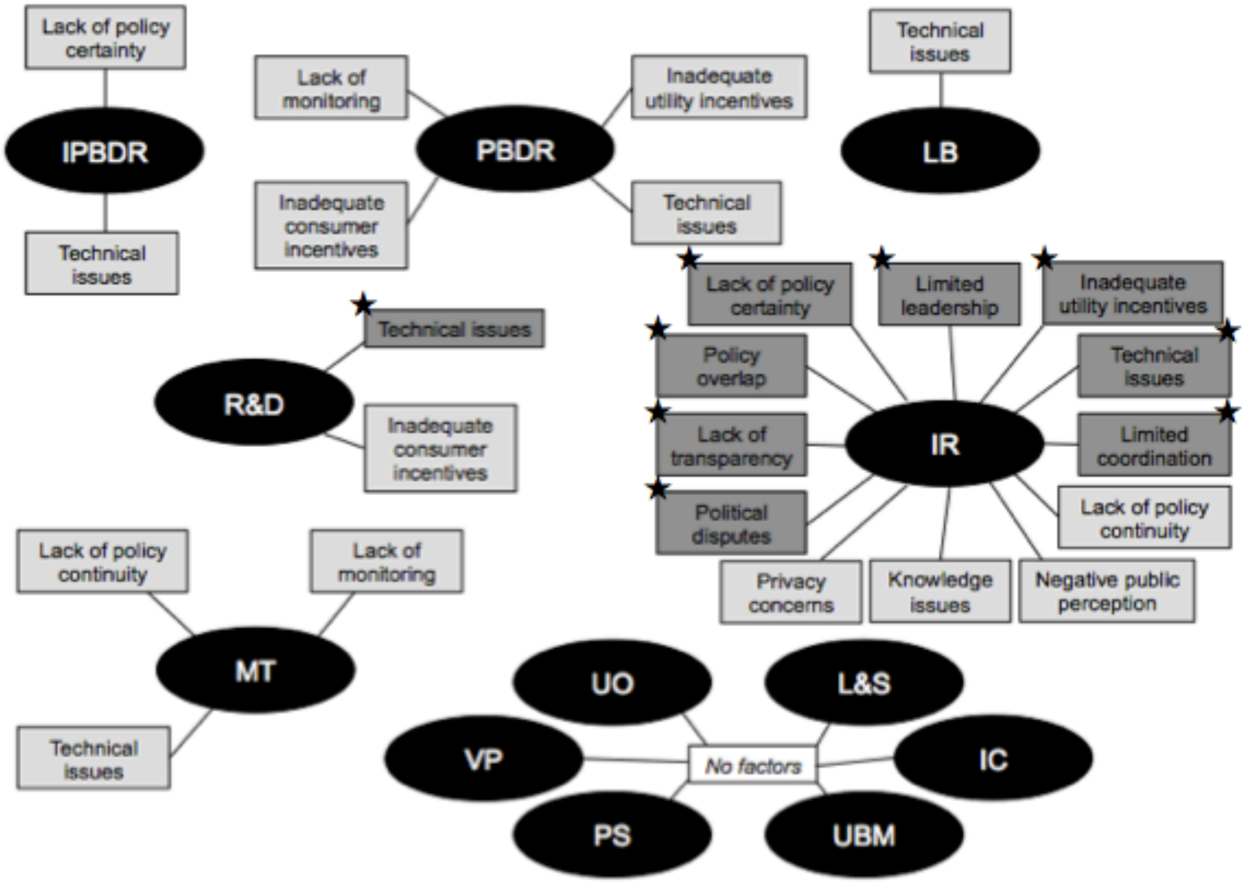


Figure 6: DSM policy mechanisms for failure

Where no factors are listed for some DSM policies, this does not mean that there are no factors that need to be considered, but that no factors were considered both frequent and highly weighted in the sample. Thus, policy makers should not ignore the factors that are frequent and not highly weighted (i.e. there is much agreement between evaluators), or infrequent and highly weighted (i.e. certain factors are important in specific contexts). However, factors that are infrequent and not highly weighted are considered unimportant and under limited government resources, these factors do not warrant the same attention.

4.3 Policy Discussions – Case Studies

Due to space, two of the twelve policies shown in figures five and six are used as case studies for in-depth discussion: utility obligations and labelling. The policies cover two quite different types of DSM policy and each is discussed in turn. However, a worked example is first given from a third policy, incentive payment-based demand response, to show how the thresholds, scales and equations were applied to the data for each of the twelve DSM policies. It is important to note that the results for each DSM policy should be viewed individually; the research does not focus on cross-comparing different DSM policies. As such, in this section, utility obligations and labelling are not compared but discussed separately. It is also important to reiterate that due to the inductive nature of the research, it does not hypothesise on policies or sectors that are not found within the 690 evaluations that make up the research sample.

Incentive payment-based demand response (IPBDR) refers to tariffs that encourage the reduction or shifting of load, particularly during peak times. From the combined analysis, regulatory frameworks was the only success factor to pass the threshold, and technical issues and a lack of policy certainty were the key failure factors.

Success Factor: Regulatory frameworks (RF)
Success Factor Frequency: 13 (above ≥5 frequency threshold)
Success Factor Weighting: 2.3 (in 1.5-2.4 'some importance' weighting group)
Policy Frequency: 62 (number of IPBDR evaluations in the sample)
Policy Weighting: 3.4 (averaged policy weighting across evaluations in sample)
Factor Combined Analysis Score: 3.4 x (13 x 2.3) / 10 = 10.5
Policy Theoretical Maximum Score: 3.4 x (62 x 5.0) / 10 = 106.4
***Combined Analysis Percentage Score:* (10.5 / 106.4) x 100% = 9.8%**

Thus, regulatory frameworks is considered the most important success factor for IPBDR, as the combined analysis score as a percentage is above the 5.0% threshold and falls into the 5.0-9.9% 'important' group (as stated previously, 'crucial' factors score ≥10.0% and 'unimportant' factors score <5.0%).

Failure Factor: Technical issues (TI)
Failure Factor Frequency: 8 (above ≥5 frequency threshold)
Failure Factor Weighting: 2.3 (in 1.5-2.4 'some importance' weighting group)
Policy Frequency: 62
Policy Weighting: 3.4
Factor Combined Analysis Score: 3.4 x (8 x 2.3) / 10 = 6.4
Policy Theoretical Maximum Score: 3.4 x (62 x 5) / 10 = 106.4
***Combined Analysis Percentage Score:* (6.4 / 106.4) x 100% = 6.0%**

The calculations show that technical issues is considered one of the two most important failure factors for IPBDR, as the combined analysis score as a percentage is above the 5.0% threshold (and falls into the 5.0-9.9% 'important' group). A similar score is produced for a lack of policy certainty (6.3%):

Failure Factor: Lack of policy certainty (LC)
Failure Factor Frequency: 8
Failure Factor Weighting: 2.4
Policy Frequency: 62
Policy Weighting: 3.4
Factor Combined Analysis Score: 3.4 x (8 x 2.4) / 10 = 6.7
Policy Theoretical Maximum Score: 3.4 x (62 x 5) / 10 = 106.4
***Combined Analysis Percentage Score:* (6.7 / 106.4) x 100% = 6.3%**

For the first case study, utility obligations (UO) usually refer to mandatory (though sometimes voluntary) obligations placed on suppliers, distributors, public entities, or building owners or users (THINK 2012). Nevertheless, the evidence base is dominated by evaluations of obligations on energy suppliers and distributors. The obligations aim to meet various policy objectives, and there are a number of ways in which the targets can be expressed. The evidence base shows that targets are commonly set in terms of energy or carbon savings with sub-targets for fuel poverty (consumers living on a low income in a home that cannot be kept warm at a reasonable cost, as defined in the UK's *Warm Homes and Energy Conservation Act 2000* and reviewed in the Hills Fuel Poverty Review 2012).

Using the same analytical process for utility obligations as shown above for incentive payment-based demand response, from the combined analysis, regulatory frameworks, legislative support, comprehensive evaluation, clear definition of roles and cost-effectiveness were the key success factors to pass the threshold, and no failure factors passed the threshold. However, failure factors were produced in the individual frequency and weighting analyses. Nevertheless, only factors that are both frequent and highly weighted are discussed in this paper.

For success factors, although utility obligations can be one of the more complex types of DSM policy to implement, it is becoming increasingly popular around the world as countries/states follow the successful experiences in the USA at a state-level and the UK. Both countries have had a long history of successfully implementing utility obligations (referred to as energy efficiency resource standards (EERS) in the USA and supplier obligations in the UK). The global evidence base shows that utility obligations require more attention being given to the regulatory and policy support factors than to the other success factor categories (financial support, stakeholder engagement and infrastructure). In the sample, the utility obligations in the UK, the USA, Italy, France and Denmark were the most evaluated countries and different evaluations of the same policies were in agreement as to the success of the policies.

In the UK, the evaluations of supplier obligations since 2002 (*Energy Efficiency Commitment*, *Carbon Emissions Reduction Target* and *Community Energy Savings Programme*) highlighted the importance of regulatory and legislative support, cost-effectiveness (for all parties concerned) and clearly defined roles for relevant parties (for example, Lees 2006, Lees 2008, Eyre *et al*. 2009, UK DECC 2011, UK DECC 2011). In the USA, similar factors are apparent for state-level EERS policies (for example, Sciortino *et al*. 2011, Neubauer *et al*. 2013, Taylor *et al*. 2012).

For the second case study, labelling (LB) refers to policies that seek to improve the communication and education of a product's energy efficiency performance. Evaluations of labelling focus on appliances, equipment and buildings in the sample. The information included on labels can vary, but the evaluations concentrated primarily on energy bill savings and carbon savings. From the combined analysis, information infrastructure was the only success factor to pass the threshold, and technical issues was the key failure factor.

For success factors, the findings reveal that label design coupled with engaging awareness campaigns is crucial (for example, Smith and Thorne 2003, Atanasiu and Constantinescu 2011). As Smith and Thorne (2003) show in the evaluation of the

ENERGY STAR and EnergyGuide labels in the USA, and Zheng *et al*. (2012) show in the evaluation of the Chinese equipment labelling schemes, label design has a reasonably limited impact on consumer perception of appliance quality or value if it is not coupled with engaging consumer awareness campaigns of the labelling schemes. Nadel *et al*. (2013) conducted an in-depth evaluation of the EnergyGuide labelling scheme ten years after Smith and Thorne (2003) and came to the same conclusion. Nadel *et al*. (2013) found that improvements to label design could be made by moving from a continuous-style graphic to a stars-based categorical comparison.

For failure factors, Smith and Thorne (2003) (USA), Zheng *et al*. (2012) (China), and Atanasiu and Constantinescu (2011) (European Union (EU)) show that technical issues in label design and communication have impacted the success of labelling policies. The USA has had a long history of energy labelling that dates back to the *Energy Policy and Conservation Act of 1975* when the mandatory EnergyGuide label was introduced for major appliances (such as refrigerators, washing machines, tumble dryers, dishwashers and air conditioners), equipment and lighting. Smith and Thorne (2003) find that, despite a reasonable familiarity of consumers with the EnergyGuide label, it appears to have had limited impact on their product choices. The same conclusion was reached by Nadal *et al*. (2013). Zheng *et al*. (2012) found that in China there was a lack of awareness of labelling enforcement due to a lack of engagement through an initial publicity campaign.

In the EU, the *Energy Labelling of Products Directive* (Directive 2010/30/EU) was implemented by member states in 2011 (replacing the previous Directive 92/75/EC energy consumption labelling scheme) to label appliances with an energy class (colour-coded letter grade: A+++, A++, A+, A, B, C, D, E, F or G), consumption and efficiency information, noise information and general appliance details. Key appliances included in the Directive are: refrigerators, washing machines, tumble dryers, dishwashers, ovens, water heaters, hot water storage tanks, air conditioners, light bulbs, televisions, cars and tyres. A related example is the EU's *Energy Performance of Buildings Directive* (2002/91/EC), which requires member states to label buildings with Energy Performance Certificates. Atanasiu and Constantinescu (2011) evaluated the Energy Performance Certificates and came to similar conclusions to Zheng *et al*. (2012), finding that the design of information and its communication were important issues affecting policy success. Thus, in summary, label design and communication (technical issues and information infrastructure) appear to be the key factors that transcend different countries and contexts.

4.4 Successful DSM Policies

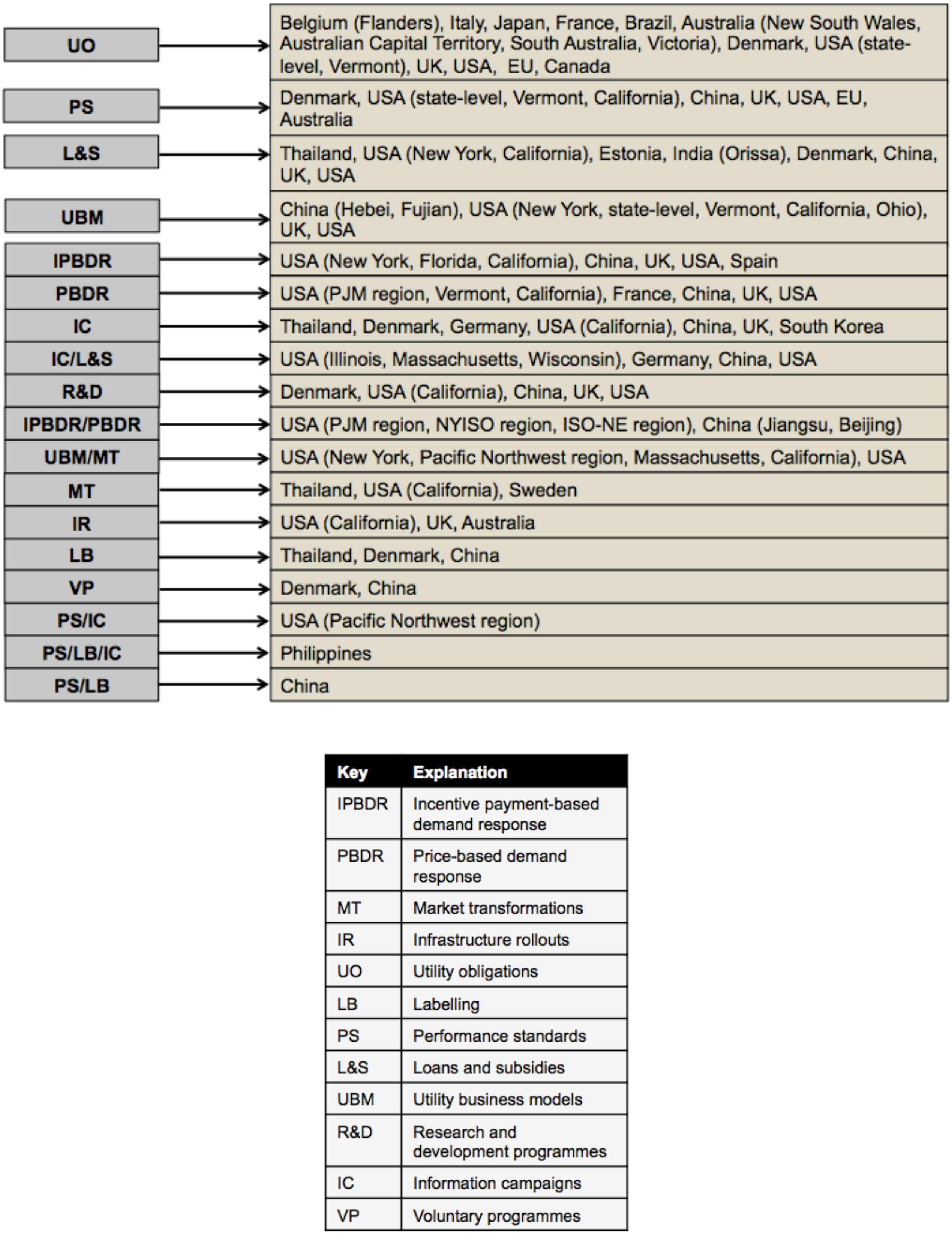
The final part of the research aimed to identify those DSM policies that have experienced more incidences of success than failure globally. The policy success weighting scale discussed previously draws parallels to the factor weighting scale, where certain words and phrases are converted into quantitative scores on the scale:

1 = 'failed', 'unsuccessful', 'ill-fated', 'ineffective' or equivalent
2 = 'less successful', 'performed poorly', 'few successes', 'less effective' or equivalent
3 = 'average performance', 'satisfactory', 'met most of the objectives' or equivalent

4 = ‘performed well’, ‘met all of the objectives’, ‘successful’, ‘effective’ or equivalent
5 = ‘highly successful’, ‘highly effective’, ‘performed beyond objectives’ or equivalent

These phrases were extracted directly from the evaluations in the production of the scale, and as such, the scale was determined inductively. Although there is arguably a bias in the judgement of the evaluators of the evaluations as to how the policy in question performed, the documents are of a high-quality (having passed the study quality assessment stage of the systematic review) and thus the expert judgement of the evaluators from conducting objective, high-quality evaluations should be considered reliable. Despite this, as discussed previously, the evaluators’ use of the same words may vary and this is an area for further methodological development.

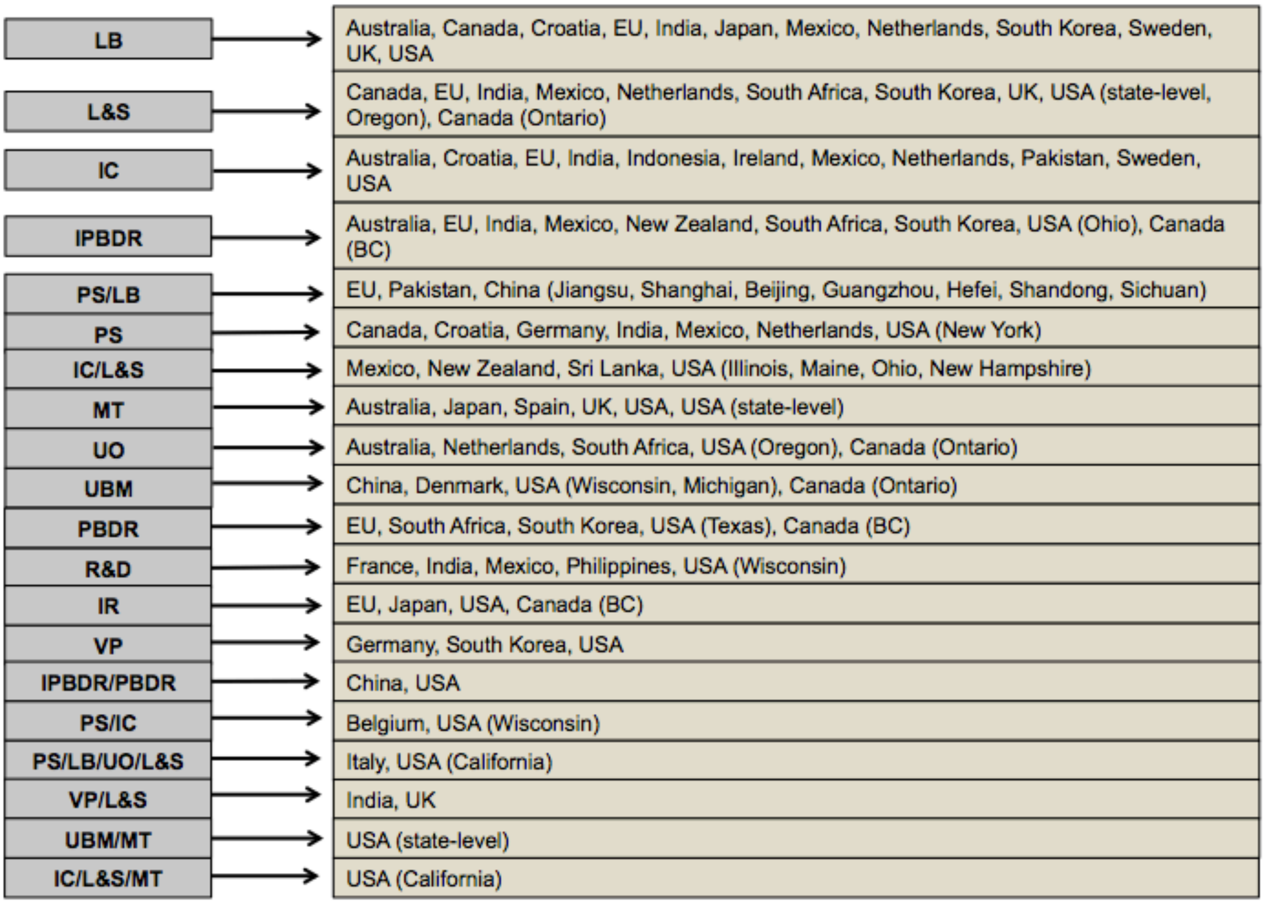
The scale was applied to each policy within each context-specific evaluation. As such, the number of countries/states that have experienced success with each of the twelve DSM policies could be identified. Policies that have a greater number of countries/states that have experienced success are considered to be more successful overall, and policies that have a greater number of countries/states that have experienced failure are considered to be more unsuccessful overall. The results are summarised in figures seven and eight and the same policy acronyms are used as per figures five and six. Where sub-national states are listed in brackets next to a country, this only refers to the specific states in question – the national level is listed separately (where relevant). A key is provided in figure seven, which is also relevant for figure eight, and where ‘/’ is used between two policies, this indicates a policy package.



UO	Belgium (Flanders), Italy, Japan, France, Brazil, Australia (New South Wales, Australian Capital Territory, South Australia, Victoria), Denmark, USA (state-level, Vermont), UK, USA, EU, Canada
PS	Denmark, USA (state-level, Vermont, California), China, UK, USA, EU, Australia
L&S	Thailand, USA (New York, California), Estonia, India (Orissa), Denmark, China, UK, USA
UBM	China (Hebei, Fujian), USA (New York, state-level, Vermont, California, Ohio), UK, USA
IPBDR	USA (New York, Florida, California), China, UK, USA, Spain
PBDR	USA (PJM region, Vermont, California), France, China, UK, USA
IC	Thailand, Denmark, Germany, USA (California), China, UK, South Korea
IC/L&S	USA (Illinois, Massachusetts, Wisconsin), Germany, China, USA
R&D	Denmark, USA (California), China, UK, USA
IPBDR/PBDR	USA (PJM region, NYISO region, ISO-NE region), China (Jiangsu, Beijing)
UBM/MT	USA (New York, Pacific Northwest region, Massachusetts, California), USA
MT	Thailand, USA (California), Sweden
IR	USA (California), UK, Australia
LB	Thailand, Denmark, China
VP	Denmark, China
PS/IC	USA (Pacific Northwest region)
PS/LB/IC	Philippines
PS/LB	China

Key	Explanation
IPBDR	Incentive payment-based demand response
PBDR	Price-based demand response
MT	Market transformations
IR	Infrastructure rollouts
UO	Utility obligations
LB	Labelling
PS	Performance standards
L&S	Loans and subsidies
UBM	Utility business models
R&D	Research and development programmes
IC	Information campaigns
VP	Voluntary programmes

Figure 7: successful DSM policies by country/state



LB	Australia, Canada, Croatia, EU, India, Japan, Mexico, Netherlands, South Korea, Sweden, UK, USA
L&S	Canada, EU, India, Mexico, Netherlands, South Africa, South Korea, UK, USA (state-level, Oregon), Canada (Ontario)
IC	Australia, Croatia, EU, India, Indonesia, Ireland, Mexico, Netherlands, Pakistan, Sweden, USA
IPBDR	Australia, EU, India, Mexico, New Zealand, South Africa, South Korea, USA (Ohio), Canada (BC)
PS/LB	EU, Pakistan, China (Jiangsu, Shanghai, Beijing, Guangzhou, Hefei, Shandong, Sichuan)
PS	Canada, Croatia, Germany, India, Mexico, Netherlands, USA (New York)
IC/L&S	Mexico, New Zealand, Sri Lanka, USA (Illinois, Maine, Ohio, New Hampshire)
MT	Australia, Japan, Spain, UK, USA, USA (state-level)
UO	Australia, Netherlands, South Africa, USA (Oregon), Canada (Ontario)
UBM	China, Denmark, USA (Wisconsin, Michigan), Canada (Ontario)
PBDR	EU, South Africa, South Korea, USA (Texas), Canada (BC)
R&D	France, India, Mexico, Philippines, USA (Wisconsin)
IR	EU, Japan, USA, Canada (BC)
VP	Germany, South Korea, USA
IPBDR/PBDR	China, USA
PS/IC	Belgium, USA (Wisconsin)
PS/LB/UO/L&S	Italy, USA (California)
VP/L&S	India, UK
UBM/MT	USA (state-level)
IC/L&S/MT	USA (California)

Figure 8: unsuccessful DSM policies by country/state

The results can be summarised as follows:

Most successful DSM policies:

- Utility obligations (16 countries/states have experienced success)
- Performance standards (9 countries/states have experienced success)
- Loans and subsidies (9 countries/states have experienced success)
- Alternative utility business models (9 countries/states have experienced success)

Least successful DSM policies:

- Labelling (12 countries/states have experienced failure)
- Loans and subsidies (11 countries/states have experienced failure)
- Information campaigns (11 countries/states have experienced failure)
- Incentive payment-based demand response (9 countries/states have experienced failure)

Here, loans and subsidies features as both a successful and an unsuccessful policy, which is reflected in its Policy Success weighting (PS_p) of 3.4 (an average policy success score). This highlights that, compared with utility obligations, performance standards and alternative utility business models, which appear to be more universally successful, loans and subsidies is more context-specific and shows examples of success in some countries/states and failure in other countries/states.

The results generally match the findings of the few studies that have been conducted in this area, such as Ürge-Vorsatz *et al*. (2007) and Harmelink *et al*. (2008). For example, Ürge-Vorsatz *et al*. (2007) found that performance standards (particularly appliance standards and building codes), loans and subsidies (particularly tax exemptions or reductions), utility obligations and labelling perform the best in terms of cost-effectiveness and carbon emissions reductions. Although this research found that loans and subsidies had an average policy success score, it was the most diverse in terms of performance across countries/states (with incidences of both success and failure). Furthermore, although the specific DSM policy level was not analysed (e.g. 'tax exemptions' within the broader category of 'loans and subsidies'), it was found that tax incentives generally performed better than subsidy policies in the sample. However, in contrast to Ürge-Vorsatz *et al*. (2007), the findings show that labelling policies have generally not performed well overall.

In addition to identifying the most and least successful DSM policies, it is similarly interesting to look at the most and least successful countries/states in implementing DSM policies, as shown below.

Most successful countries/states:

- California (USA) (10 policy types successfully implemented)
- China (10 policy types successfully implemented)
- UK (9 policy types successfully implemented)
- USA (9 policy types successfully implemented)

Least successful countries/states:

- European Union (EU) (7 policy types unsuccessfully implemented)
- India (7 policy types unsuccessfully implemented)
- Mexico (7 policy types unsuccessfully implemented)
- USA (6 policy types unsuccessfully implemented)

California, China, the USA and the UK have experienced success with the full range of DSM policy categories from demand response policies and large-scale research and development programmes to performance standards and alternative utility business models. In contrast, the evidence base shows that the EU, India, Mexico and the USA have experienced policy failure with a large range of DSM policy categories. In the case of the USA, its experience with DSM policy is the greatest of any country/state, which explains why it has both a high number of successful and unsuccessful policies.

It is important to note that any of the DSM policies examined can be successfully implemented in any of the countries/states in the sample if the identified success and failure factors are taken into account. The majority of policies and countries/states in the sample showed incidences of both success and failure. Furthermore, policy evaluations conducted since 2014 will not have been included, as this was when the data collection was completed. However, the methodological approach has been designed so that it can be readily updated as new evidence is produced.

5. Conclusion

Much of the research that has been undertaken in the demand-side policy field has focused on policy impacts rather than the mechanisms behind policy success and failure. This paper contributed to filling this knowledge gap by presenting the results of a four-year project to explore how and why demand-side policies succeed or fail, and what policies have been successful. The research covered twelve different types of demand-side management (DSM) policy, which includes policies for energy efficiency, demand response, on-site generation and on-site storage.

The research provided the first systematic review of the global evidence base on DSM policy, and specifically employed the realist synthesis type of systematic review, which has received limited attention in the energy policy field compared with other disciplines. The research identified the key factors for success and failure in 690 DSM policy evaluations, which were included within 102 high-quality documents that covered 30 countries and 36 sub-national states across six continents. Previous related studies have concentrated on specific contexts, such as a single DSM policy or country/region, rather than identifying key factors that are common across contexts for different types of DSM policy. Limitations of previous studies include small sample sizes, a lack of focus on identifying failure factors (which this research has found to not simply be the absence of identified success factors), limited attention to the importance of identified factors when compared with other factors for a given DSM policy, and limited use of rigorous systematic review techniques.

In addition to its empirical contribution to the DSM field, the research proposed new methodological techniques to determine the key success and failure factors overall, the factors for each of the twelve DSM policies examined, and successful DSM policies. The approach used factor frequency and weighting analyses and proposed an equation to combine the analyses in order to identify factors that were both frequent and highly weighted in the sample. The analytical techniques can be readily applied to other areas of energy and climate policy.

Across DSM policy types and countries / sub-national states, the overall findings show that regulatory frameworks and appropriate incentives are the most important success factors and a lack of monitoring (for evaluation) and technical issues (primarily programme management issues) are the most important failure factors. Thus, above all other factors, governments need to provide the required regulatory frameworks and appropriate incentives for demand-side policies to succeed, but in parallel, the policies need to be monitored throughout the lifecycle of the policy period and enough resources need to be dedicated to the proper administration and evaluation of the policy. The more insightful findings from the research are the results broken down by DSM policy type, which are summarised in figures five and six, and these form the main policy recommendations for the future design of demand-side policies. These findings are based on the development and synthesis of a robust DSM policy evidence base using rigorous systematic review techniques, which has identified factors that are defendable across different countries and contexts.

Further research should analyse the results at the specific policy level (such as appliance standards rather than performance standards – e.g. Houde and Spurlock 2016), in order to provide a lower level of analysis to inform policy decisions. Furthermore, further research could look to adapt the new methodological techniques for application in other areas of energy and environmental policy to provide further testing and methodological validation.

Acknowledgements

This research was made possible by UK EPSRC support, grant number EP/H009612/1. I would like to thank Professor Mark Barrett (UCL), Professor Paul Ekins (UCL), Andrew Smith (UCL), Professor David Shipworth (UCL), Michelle Shipworth (UCL), Professor Steven Sorrell (University of Sussex), Professor Diana Ürge-Vorsatz (Central European University), and the anonymous peer reviewers for their comments on the research.

Dataset

The list of documents that make up the systematic review research sample can be found in Warren (2015).

Bibliography

- Agnew, K., R. Burke, and P. Ham-su. 2009. "Participation of demand response resources in ISO New England's Ancillary Service Markets." *International Energy Program Evaluation Conference 2009*.
- Atanasiu, B., and T. Constantinescu. 2011. "A comparative analysis of the energy performance certificates schemes within the European Union: Implementing options and policy recommendations." *ECEEE Summer Study 2013 Proceedings*. European Council for an Energy-Efficient Economy (ECEEE).
- Boudet, H., N.M. Ardoin, J. Flora, K.C. Armel, M. Desai, and T.N. Robinson. 2016. "Effects of a behaviour change intervention for Girl Scouts on child and parent energy-saving behaviours." *Nature Energy* 1.
- Cappers, P., C.A. Goldman, and D. Kathan. 2009. "Demand Response in US Electricity Markets: Empirical Evidence." Report number LBNL-2124E. Ernest Orlando Lawrence Berkeley National Laboratory. USA.
- Carbon Emissions Reduction Target (CERT). 2008. Department of Energy and Climate Change (DECC). UK.
- Community Energy Saving Programme (CESP). 2008. Department of Energy and Climate Change (DECC). UK.
- Dixon-Woods, M., S. Agarwal, D. Jones, B. Young, and A. Sutton. 2005. "Synthesising qualitative and quantitative evidence: a review of possible methods." *Journal of Health Services Research Policy* 10 (1): 45-53.
- Energy Efficiency Commitment 1 (EEC 1). 2002. Department of Energy and Climate Change (DECC). UK.
- Energy Efficiency Commitment 1 (EEC 2). 2005. Department of Energy and Climate Change (DECC). UK.
- Energy Policy Act of 2005. Washington DC: National Congress. USA.
- European Council Directive 2002/91/EC on energy performance of buildings.
- European Council Directive 92/75/EEC on energy labelling of products.
- Eyre, N., M. Pavan, and L. Bodineau. 2009. "Energy company obligations to save energy in Italy, the UK and France: what have we learnt?" *ECEEE Summer Study 2013 Proceedings*. European Council for an Energy-Efficient Economy (ECEEE).
- Gandhi, R., C.R. Knittel, P. Pedro, and C. Wolfram. 2016. "Running randomized field experiments for

energy efficiency programs: A Practitioner's Guide." *Economics of Energy & Environmental Policy* 5 (2).

- Hahn, R., R. Metcalfe. 2016. "The impact of behavioral science experiments on energy policy. *Economics of Energy & Environmental Policy* 5 (2).
- Harmelink, M., L. Nilsson, and R. Harmsen. 2008. "Theory-based policy evaluation of 20 energy efficiency instruments." *Energy Efficiency* 1: 131-148.
- Hills, J. 2012. "Getting the measure of fuel poverty: final report of the Fuel Poverty Review." CASE report 72. *Centre for Analysis of Social Exclusion (CASE)*. March 2012
- Houde, S. and C.A. Spurlock. 2016. "Minimum energy efficiency standards for appliances: old and new economic rationales." *Economics of Energy & Environmental Policy* 5 (2).
- Lees, E. 2006. "Evaluation of the Energy Efficiency Commitment 2002-2005." Report to DEFRA. Prepared by Eoin Lees Energy. 28th February 2006. UK
- Lees, E. 2008. "Evaluation of the Energy Efficiency Commitment 2005-2008." Report to DEFRA. Prepared by Eoin Lees Energy. 28th February 2006. UK.
- McConnell, A. (2010) "Policy success, policy failure and grey areas in-between." *Journal of Public Policy* 30: 345-362.
- Multi-Criteria Decision Making (MCDM) analysis interview with the UK Demand Response Association (UK DRA). 06/06/2014. London. UK.
- Nadel, S., J. Amann, S. Hayes, S. Bin, R. Young, E. Mackres, H. Misuriello, and S. Watson. 2013. "An Introduction to US Policies to Improve Building Efficiency." *American Council for an Energy-Efficient Economy (ACEEE)*. Research Report A134. July 2013.
- Neubauer, M., B. Foster, N. Elliott, D. White, and R. Hornby. 2013. "Ohio's Energy Efficient Resource Standard: Impacts on the Ohio Wholesale Electricity Market and Benefits to the State." *American Council for an Energy-Efficient Economy (ACEEE)*. Research Report E138. April 2013.
- Nicolson, M., G.M. Huebner, and D. Shipworth. 2017. "Are consumers willing to switch to smart time of use electricity tariffs? The importance of loss-aversion and electric vehicle ownership." *Energy Research & Social Science* 23: 82-96.
- Nicolson, M., G.M. Huebner, and D. Shipworth. 2017. "Tailored emails prompt electric vehicle owners to engage with tariff switching information." *Nature Energy* 2.
- Pawson, R. 2002. "Evidence-based policy: in search of a method." *Evaluation* 8 (2): 157-181.
- Pawson, R. 2002. "Evidence-based policy: the promise of ‘Realist Synthesis." *Evaluation* 8 (3): 340-35.
- Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: SAGE.
- Petticrew, M., and H. Roberts. 2006. *Systematic Reviews in the Social Sciences*. Blackwell Publishing. Oxford. UK
- Patricia J. Rogers, A.P Tracy, A. Huebner, and T.A. Hacsi. 2000. "Program Theory Evaluation: Practice, Promise, and Problems." *New Directions in Evaluation* 87 (Fall)
- Rosenow, J., C. Leguijt, Z. Pato, N. Eyre, and T. Fawcett. 2016. "An ex-ante evaluation of the EU Energy Efficiency Directive: Article 7." *Economics of Energy & Environmental Policy* 5 (2).
- Rossi, P., M. Lisey, and H. Freeman. 2004. *Evaluation: a systematic approach*. 7th edition. Thousand Oaks: SAGE.
- Schultz, P.W., M. Estrada, J. Schmitt, R. Sokoloski, and N. Silva-Send. 2015. “Using in-home displays to provide smart meter feedback about household electricity consumption: a randomised control trial comparing kilowatts, cost and social norms.” *Energy* 90 (1): 351-358.
- Sciortino, M., S. Nowak, P. Witte, D. York, and M. Kushler. 2011. "Energy Efficiency Resource Standards: A Progress Report on State Experience." *American Council for an Energy-Efficient Economy (ACEEE)*. Research Report U112. June 2011.
- Siano, P. 2014. “Demand response and smart grids – a survey”. *Renewable and Sustainable Energy Reviews*. 30: 461-478.
- Smith, S., and J. Thorne, J. 2003. "An evaluation of the EnergyGuide Label: what we learned." *International Energy Program Evaluation Conference 2003*.
- Snilstveit, B., S. Oliver, and M. Vojtkova. 2012. “Narrative approaches to systematic review and synthesis of evidence for international development policy and practice.” *Journal of Development Effectiveness*. 4 (3): 409-429.
- Sorrell, S. 2007. “Improving the evidence base for energy policy: the role of systematic reviews.” *Energy Policy*. 35: 1858-1871.

- Stern, F. and D. Vantzis. 2014. “Protocols for Evaluating Energy Efficiency – Both Sides of the Atlantic.” *International Energy Policy and Programme Evaluation Conference*. 9-11th September 2014. Berlin. Germany.
- Taylor, B., D. Trombley, and J. Reinaud. 2012. "Energy Efficiency Resource Acquisition Program Models in North America." *American Council for an Energy-Efficient Economy (ACEEE)*. Research Report IE126. November 2012.
- THINK 2012. “How to refurbish all buildings by 2050.” Topic 7. Final report. June 2012.
- UK Civil Service. 2014. “What is a rapid evidence assessment?” UK Civil Service website: http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/what-is
- UK Department of Energy and Climate Change (DECC). 2011. "Evaluation of the Low Carbon Buildings Programme." Research Report. August 2011. DECC.
- UK Department of Energy and Climate Change (DECC). 2011. "Evaluation of the Community Energy Saving Programme." Research Report. October 2011. DECC.
- UK Department of Energy and Climate Change (DECC). 2011. "Energy supplier obligation policies: evaluation synthesis." Research Report. October 2011. DECC.
- Ürge-Vorsatz, D., S. Koeppel, and S. Mirasgedis, S. 2007. “Appraisal of policy instruments for reducing buildings’ CO_2 emissions.” *Building Research & Information* 35 (4): 458-477.
- Warm Homes and Energy Conservation Act of 2000. London: HMSO. UK
- Warren, P. 2014. “A review of demand-side management policy in the UK.” *Renewable and Sustainable Energy Reviews* 29: 941-951.
- Warren, P. 2014. “The use of systematic reviews to analyse demand-side management policy”. *Energy Efficiency* 7: 417-427.
- Warren, P. 2015. *Demand-side management policy: mechanisms for success and failure*. PhD thesis. University College London (UCL). UK.
- Warren, P. 2017. “Transferability of demand-side policies between countries.” *Energy Policy* 109: 757-766. DOI: 10.1016/j.enpol.2017.07.032.
- Warren, P. 2018. “Evidence reviews in energy and climate policy.” *Evidence & Policy.* DOI: 10.1332/174426418x15193815413516.
- Warren, P. 2018. “Demand-side policy: global evidence base and implementation patterns.” *Energy & Environment*. DOI: 10.1177/0958305x18758486.
- Wong, G., T. Greenhalgh, G. Westhorp, J. Buckingham, and R. Pawson. 2013. “RAMESES publication standards: realist synthesis.” *BMC Medicine*. 11 (21): 1-14.
- Zarnikau, J.W. 2010. "Demand participation in the restructured Electric Reliability Council of Texas market." *Energy* 35 (4): 1536-1543.
- Zheng, N., N. Zhou, C. Fino-Chen, and D. Fridley. 2012. "Evaluation of local enforcement of energy efficiency standards and labeling program in China." *International Energy Program Evaluation Conference 2012*.