US protocol to evaluate behavioural measures in the residential sector

Background and objectives of the UMP protocol on residential behaviour

Context and objectives

The U.S. Department of Energy launched in 2012 the Uniform Methods Project (UMP)[1] to establish a set of model protocols for determining energy and demand savings that result from energy efficiency measures and programmes implemented through state and utility energy efficiency programmes in the US. In many states, energy utilities are required or incentivized to implement energy efficiency programmes. To recover their cost of implementing these programmes, utilities shall report evaluations to the Public Utility Commission of their state.

The protocols prepared by the UMP provide detailed descriptions of the commonly accepted evaluation methods to help ensure that similar programs are measured in the same way (harmonisation purpose). Each chapter of the UMP deals with a given type of energy end-use and sector, complemented with cross-cutting chapters. Each chapter has been written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment.

The UMP protocols are meant for utilities, programme administrators, public utility commissions, evaluators, and other stakeholders for both programme planning and evaluation.

Scope of the protocol on 'Residential Behavior'

This protocol (Stewart and Todd 2020) deals with 'residential behavior-based (BB) programmes' that include the following types of behavioural interventions (possibly combined):

- providing end-users with feedback about their real-time or historical energy consumption;
- reframing of consumption information in different ways;
- supplying energy efficiency education and tips;
- rewarding end-users for reducing their energy use;
- comparing end-users to their peers; and
- establishing games, tournaments, and competitions.

[1] For more details, see: https://www.energy.gov/eere/uniform-methods-project-determining-energy-efficiency-program-savings
All UMP protocols can be found here: https://www.energy.gov/eere/uniform-methods-project-determining-energy-efficiency-savings-specific-measures

These programmes, deployed on large scale since 2008, often target multiple energy end-uses. Their design may vary broadly. Nevertheless, the protocol identified common features, including:

- the use of randomized experiments: eligible end-users are randomly assigned to treatment or control groups;
- large scale: each programme targets thousands of end-users;
- providing end-users with tailored information (analyses of their historical consumption, energy savings tips, and energy efficiency comparisons to neighbouring homes) or offering incentives for saving energy.

The 2020 update of the protocol also mentioned the development of alternative programme designs, implemented at smaller scale with new communication channels (e.g. social media and text messaging), or with new types of intervention (e.g. social media competitions to reduce consumption). The protocol explained that these new types of programmes may raise new evaluation issues that cannot necessarily be addressed with the methods included in the 2020 protocol. In particular, the types of programmes covered by the protocol can usually be evaluated with randomized experiments. Whereas these new types of programmes may be evaluated with quasi-experimental methods (due to the smaller scale) that may involve more uncertainties.

Evaluation methods

Recommended methods: RCTs and REDs

The UMP protocol recommends the use of **Randomized Control Trials (RCTs)** or **Randomized Encouragement Designs (REDs)**, presented as the gold standard for evaluating behavioural measures, based on comparisons with quasi-experimental methods and observational studies (see e.g. Allcott 2011; Spurlock et al. 2016).

The protocol recognizes that implementing RCTs or REDs may not always be feasible, and suggests that in these cases, the second-best option is to use **quasi-experimental methods**. These methods require strong assumptions, and when these assumptions are not met, the results may be biased. The uncertainty range is then difficult to assess. Due to this, quasi-experimental methods are not described in the protocol that refers to other sources about them (e.g. SEE Action 2012, Cappers et al. 2013).

RCTs, REDs and quasi-experimental methods are assessing energy savings by comparing changes in energy consumption over time between two groups: a **comparison group** (providing the baseline) and a **treatment group** (where the effects of the behavioural interventions can be observed).

The main difference between RCTs, REDs and quasi-experimental methods is that with RCTs and REDs the comparison group is **defined randomly** with sampling criteria that can be controlled. The comparison group is then called control group. The random assignment of end-users to the control group (or treatment group) ensures that receiving the treatment is uncorrelated with end-users' pretreatment energy use. Therefore, any difference in energy use between the control and treatment groups can be attributed to the behaviour interventions evaluated (internal validity).

Whereas the comparison group in quasi-experimental methods is not defined randomly, and usually only part of the sampling criteria can be controlled. Hence the need for quasi-experimental methods to use complementary approaches (e.g. matching) to reduce the risk of bias (especially selection bias).

There can also be several treatment groups, when the evaluation aims at comparing different types of interventions.

The main difference between RCTs and REDs is that, in RCTs, the treatment group receives a programme treatment (i.e. behavioural intervention(s)). Whereas in REDs, the treatment group receives encouragement[2] to participate in the programme. Therefore, the result assessed in **RCTs** is the **effect of the behavioural intervention(s)**, whereas the result assessed in **REDs** is the **effect of the encouragement to participate in the programme**[3]. Both results are usually measured in terms of energy (or demand) savings.

Choosing between RCT or RED depends on the type of programme, expected number of participants and how the programme can accommodate the constraints of the evaluation method. A particular constraint for an RCT is that it might require delaying or denying participation for the end-users selected for the control group. REDs can then be preferred to allow all end-users to participate.

In any case, using RCTs or REDs requires **upfront planning**. This implies that evaluation of energy savings must be included in the design and overall planning of the programme.

The methods recommended in the protocol correspond to Option C of IPMVP (International Performance Measurement and Verification Protocol), i.e. statistical analysis of data from utility meters for whole dwellings to estimate savings. This option is usually recommended for interventions with expected savings that are large relative to the whole dwelling's energy consumption. This is why the protocol highlights the importance for the samples (control and treatment groups) to be large enough to ensure statistical significance. The protocol thus recommends regression analysis of

[2] Typical examples of encouragement include direct paper mail or email informing end-users about the opportunity to participate in a programme.
[3] The result obtained from REDs is also called Local Average Treatment Effect (LATE).

energy consumption and statistical power analysis to determine the sample size necessary to detect the expected savings (for more details, see section 4.2.1 of the protocol).

General process to implement RCTs or REDs

The protocol summarises the following key steps to implement RCTs or REDs:

- Identify the **study population** (cf. eligibility criteria and targeting of the programme)
- Identify the **treatment effect** (usually energy or demand savings) and the **measurement approach** (e.g. measurement period, metering frequency)
- Determine **sample sizes**, that depends on the type of programme (e.g. opt-in or opt-out) and evaluation method (RCT or RED), magnitude of expected savings, variance of energy consumption, and required confidence level (or statistical significance) (for more details, see section 4.2.1 of the protocol).
- **Randomly assign end-users to control and treatment groups**: the protocol recommends about this point that the random assignment be done by a qualified independent third party, and that the assignment procedure be transparent and well documented (for more details, see section 4.2.2 of the protocol).
- Verify **equivalence** of samples (i.e. no significant difference in average and distribution of pretreatment energy consumption between the control and treatment groups) (for more details, see section 4.2.3 of the protocol).
- Administer the **treatment** (either once or repeatedly, depending on the programme and types of interventions), with a treatment period long enough for effects to be observable.
- **Collect data** from all groups, including pre-treatment energy consumption data whenever possible (for more details, see section 4.3 of the protocol).
- **Estimate savings**, usually with a difference in differences (DiD) method comparing the changes over time in the control and treatment groups (whether or not end-users actually complied with the treatment in the treatment group). The protocol also recommends using panel regression analysis (for more details, see section 4.4).

For more details, see the section 3.4 of the protocol (Stewart and Todd, 2020) that describes the use of RCT or RED to evaluate typical behavioural programmes implemented by utilities in the US:

- **opt-out** programmes: where end-users may ask the utility to stop receiving support or information that was administered without prior agreement from the end-users (e.g. home energy reports sent to the whole treatment group);

- **opt-in** programmes: where end-users enrol in the programme to receive support or information (such as web-based home audit or energy consumption tools, online courses about energy rates and home efficiency; or in-home displays)[4].

Lifetime and persistence of energy savings

The protocol mentions a usual assumption of savings lifetime of 1 year. It is presented as a conservative assumption, considering uncertainty that effects (i.e. changes in behaviours) last after the interventions end, and the lack of evidence that would show longer lifetime.

Researchers have addressed the lack of evidence by developing frameworks for estimating savings persistence (see e.g. Khawaja and Stewart 2014; Jenkins et al. 2017). These frameworks account for repeated, multiyear program treatments and the gradual decay of energy savings from behavioural interventions, usually distinguishing three key concepts:

- **intervention period**: for how long the behavioural interventions are implemented
- savings **persistence**: for how long energy continues to be saved by the end-users
- savings **decay**: rate of decrease in energy saving effect over time

When dealing with new annual savings, **double counting** should be avoided by taking into account that when interventions are implemented in multiple years, some annual savings may be attributable to interventions done in previous years.

Another common issue is **attrition or withdrawal**. For example, end-users may move residences. Similarly, the longer the period monitored, the higher the chances for changes in the composition of households, etc.

For more details about evaluating savings persistence and decay, see section 4.6 of the protocol.

[4] For opt-in programmes, the random assignment of end-users to control or treatment groups happens once end-users have enrolled on the programme. End-users who do not enroll on the programme are not part of the study population. The estimate of savings will still have internal validity, but not necessarily external validity (as it will not apply to end-users who did not enroll on the programme). When using RCT, this also means that programme administrators will deny or delay the participation of part of the end-users who enrolled on the programme (the ones assigned to the control group). This can be avoided by using RED instead of RCT.

Main results or findings

Based on available reviews (e.g. Allcott 2011; David 2011; Rosenberg et al. 2013), the UMP protocol indicates that **energy savings** from behaviour-based programmes are usually a small percentage of energy use, **typically less than 3%.**

The protocol does not recommend indicative values about savings lifetime or decline over time. However, it pointed out the **development of studies in this field** in the 2010's, growing the evidence base that behavioural measures can have effects lasting after the interventions end, with energy savings sometimes persisting for several years. These results may vary significantly depending on the types and duration of interventions. Moreover, there are still discussions about the **causes of savings persistence**: whether this would come from the adoption of long-lasting savings habits or from investments (e.g. replacing appliances with more efficient ones, or renovating the building envelope). This reiterates the importance of considering **possible overlaps and double counting** with policy measures promoting energy efficiency investments in the residential sector.

With respect to evaluation practices, a key point is defining the **sample size**. The protocol mentions that, as behavioural measures have often expected savings smaller than 3%, this may require **thousands of end-users** in the treatment and control groups. The samples need to be even larger when using REDs to have enough statistical power to detect the expected effect.

When **comparing RCT and RED**, the protocol notes that RCTs work well with opt-out programmes (e.g. home energy reports) and with opt-in programmes, for which participation can be delayed (e.g. using a waiting list) or denied (e.g. when the number of end-users that can be supported is limited). For situations in which delaying or denying participation is impossible, costly or unwanted, REDs may be more appropriate. REDs can accommodate all end-users, but require larger analysis samples and are more complex to perform and explain (cf. two analysis steps to get an estimate of the saving effect on energy use).

Moreover, the protocol argues that limiting the participation to certain end-users (as done in RCTs) provides benefits compensating the drawbacks: advancing the knowledge about saving effects of behavioural interventions, and providing evidence to optimize the allocation of resources to achieve energy savings.

Another key practical point is **data collection**. The protocol recommends that the collection of energy consumption data be made for at least the duration of the interventions, and at least for one full year of historical energy use data (i.e. the 12

months before the interventions start). The latter is to ensure that the baseline reflects seasonal energy use effects. It is also very useful to control for fixed effects.

The protocol also discusses the data frequency, highlighting that hourly energy consumption data may enable the estimation of precise savings during peak hours, which may be increasingly important with higher electrification of end-uses and share of RES in the electricity system.

Replication potential in the context of Article 8 EED (about half a page)

Stewart and Todd discuss what **conditions** should be met for the evaluation protocol to be applicable:

- An **appropriately sized analysis sample** can be constructed: applying this protocol requires that the analysis sample be large enough to detect the expected savings with high probability. As most behavioural programmes result in small percentage savings, a large sample size (often in the thousands or tens of thousands of end-users) is essential to detect savings.
- **Participants** ('treated end-users') **can be identified** and **reliable energy use data** is available for a sample of end-users: the protocol cannot be used when it is not possible to monitor who is (non-)participant. Moreover, monitoring & evaluation should plan in advance the data collection about energy use, that should be based on metered data.
- It must be possible to **isolate the treatment effect** when measuring savings (e.g. restricted programme participation or marketing to randomly selected end-users): this is the purpose of using RCT or similar experimental method. The evaluation method should consider the expected result from the behavioural measures (reducing energy consumption, reducing peak demand, both), and what periods should be monitored.
- The evaluation methods described in the protocol imply that **evaluation should be planned in advance**, with evaluators involved early in the programme design.

It should be noted that the methods described in the protocol can apply whatever the type of intervention and communication channel used, as long as the conditions above are met.

The protocol cannot be used to evaluate energy savings from a mass-media or social media campaign promoting energy saving behaviours, as it would not be possible to isolate a control group, nor to identify who receives the messages.

The protocol also highlights the need to consider possible overlaps and **double counting** with other programmes. Behavioural measures can indeed increase the participation to other programmes (e.g. grant scheme to replace heating systems)[5]. When reporting energy savings to Article 8 EED, Member States should avoid this type of double counting.

Another key point for the use of the methods described in the US protocol is '**external validity**', defined as follows: "*Savings estimates are externally valid if evaluators can apply them to different populations or time periods from those studied*". Stewart and Todd warn that caution is needed when applying savings estimates for one population to another, or to the same population at a later time. Differences in population characteristics, weather, other external factors (e.g. energy prices) or overlaps with other policy measures can change the energy savings that can be attributed to the programme under evaluation.

References

The UMP protocol on Residential Behaviour:

Stewart, J. and Todd, A. (2020). Chapter 17: Residential Behavior Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. October 2020.

Complementary references:

Allcott, H. (2011). Social Norms and Energy Conservation. *Journal of Public Economics*, 95(2), 1082-1095.

Allcott, H., and Rodgers, T. (2014). The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. *American Economic Review*, 104 (10), 3003-3037.

Cappers, P., Todd, A., Perry, M., Neenan, B., Boisvert, R. (2013). Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines. LBNL (Lawrence Berkeley National Laboratory) and EPRI (Electric Power Research Institute), LBNL-6301E.

Davis, M. (2011). Behavior and Energy Savings: Evidence from a Series of Experimental Interventions. Environmental Defense Fund Report.

EPRI (2010). Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols. Electric Power Research Institute: Palo Alto, CA.

[5] This effect is named 'efficiency program uplift' in the US protocol. For more details, see section 4.5 of the US protocol.

Ignelzi, P., Peters, J., Dethman, L., Randazzo, K., Lutzenhiser, L. (2013). Paving the Way for a Richer Mix of Residential Behavior Programs. Prepared for Enernoc Utility Solutions. CALMAC Study SCE0334.01.

Jenkins, C., Weaver, T., Olig, C., Patterson, O. and Brightwell, D. (2017). Accounting for Behavioral Persistence-A Protocol and a Call for Discussion. Proceedings of IEPEC (International Energy Program Evaluation Conference) 2017.

Khawaja, M. S. and Stewart, J. (2014). Long-Run Savings and Cost Effectiveness of Home Energy Report Programs. Cadmus Research Report.

Mazur-Strommen, S., Farley, K. (2013). ACEEE Field Guide to Utility-Run Behavior Programs. American Council for an Energy Efficient Economy, Report Number B132

Minnesota Department of Commerce, Division of Energy Resources (2015). Energy Efficiency Behavioral Programs: Literature Review, Benchmarking Analysis, and Evaluation Guidelines. Prepared by Illume Advising with subcontractors Indica Consulting and Pr. Edward Vine.

Rosenberg, M., Kennedy Agnew, G., Gaffney, K. (2013). Causality, Sustainability, and Scalability – What We Still Do and Do Not Know about the Impacts of Comparative Feedback Programs. Proceedings of IEPEC (International Energy Program Evaluation Conference) 2013.

SEE Action (2012). Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations. State and Local Energy Efficiency Action Network. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, LBNL.

Spurlock, A., Cappers, P., Jin, L., Todd, A., Baylis, P. (2016). Go for the Silver? Evidence from field studies quantifying the difference in evaluation results between "gold standard" randomized controlled trial methods versus quasi-experimental methods. Proceedings of the ACEEE 2016 Summer Study on Energy Efficiency in Buildings.

Thomas, A. (2022). US experience with measuring energy savings from behavioural programmes. Presentation at the streamSAVE dialogue meeting, 15 November 2022.