Are we heading towards a replicability crisis in energy efficiency research? A toolkit for improving the quality, transparency and replicability of energy efficiency impact evaluations

Gesche M. Huebner[1]†
Email: g.huebner@ucl.ac.uk

Moira L. Nicolson†
Email: m.nicolson.11@ucl.ac.uk

Michael J. Fell†
Email: michael.fell@ucl.ac.uk

Harry Kennard†
Email: harry.kennard.13@ucl.ac.uk

Simon Elam†
Email: s.elam@ucl.ac.uk

Clare Hanmer†
Email: clare.hanmer.15@ucl.ac.uk

Charlotte Johnson†
Email: c.johnson@ucl.ac.uk

David Shipworth†
Email: d.shipworth@ucl.ac.uk

†UCL Energy Institute
14 Upper Woburn Place
GB – London, WC1H

Abstract

Several high-profile replication failures have called into question the reproducibility of results in medicine, neuroscience, genetics, psychology and economics (Camerer et al. 2016). A paper published in Science found that just one third of psychology studies could be replicated when the study was run for a second time (OSC 2015). To our knowledge, there have been no attempted replications of energy efficiency studies; so can we be confident that the estimated energy savings from policy initiatives like the European roll out of smart meters will be realised? Or that electric vehicles will reduce carbon emissions by predicted levels? Or is energy heading towards its own reproducibility crisis? Researchers call for the increased use of randomised control trials (RCTs) to evaluate energy efficiency policy and the introduction of protocols or guidelines for conducting experiments (Vine et al. 2014; Frederiks et al. 2016). However, no guidelines for increasing reproducibility have been proposed. Moreover, RCTs are just one method for causal analysis and RCTs cannot answer all important causal questions. This paper will outline research methods for improved impact assessment of energy efficiency policy, including RCTs, but also quasi-experiments and systematic reviews that go beyond the conclusions of single experiments. It will then present tools for increasing replicability: pre-registration of trials; pre-analysis plans; reporting standards; synthesis tools and; publication of datasets with computer code in data repositories. Based on work by our research group at the UCL Energy Institute, we recognize that not all of these tools (mostly from medical trials) provide 'off-the-shelf' models for energy efficiency evaluations, and so consider adaptations for energy research. Our aim is to stimulate discussion and get feedback from the research community at ECEEE so the toolkit can be developed and potentially adopted more widely.

Introduction

Recent years have seen the emergence of a replication crisis in various academic fields such as medicine, neuroscience, genetics, psychology and economics (Camerer et al. 2016) where attempts to replicate previous results have failed extensively. Psychology was probably the hardest-hit field with a recent publication showing that only about one third of seminal psychology studies published in high ranking journals could be replicated when the study was run for a second time (Open Science Collaboration et al. 2015). In addition, even when results could be replicated, i.e. a significant effect in the same direction was found as in the original study, the

[1] Huebner, Nicolson, and Fell are joint first authors of the paper and have contributed equally.

effect size was often much lower than in the original study. Here, replication refers to the process by which the original or independent researcher(s) carry out the same study for the specific purpose of testing whether the second study reaches the same results as the original, generally using the same methods but collecting new data.[2] This is to be distinguished from reproducibility where the same computer code is run on the original data to reproduce exact numerical.

Energy research has so far been spared from any replication or reproducibility crisis; however, this might just be down to the fact that there are hardly any attempts at replication or reproduction within the energy sector. Therefore, how do we know that our findings are robust and valid? Many of the findings we trust today (e.g. the energy savings from smart meters, impact of building energy ratings and of household energy efficiency improvements) remain unchecked. This is concerning because the replicability/reproducibility of energy research could be lower than in other disciplines because no energy journals require authors to provide the accompanying data and step-by-step information on how the data was collected and analysed to even permit a replication or reproduction to be conducted. By comparison, at least three of the top tier economics journals (*American Economic Review*, *Econometrica*, *Journal of Applied Economics*) will not permit authors to submit a manuscript for peer-review without also submitting the dataset(s) and computer programs necessary to replicate the results in the paper. Given that energy efficiency programmes have not been subject to the same rigorous, empirical evaluation methods that are considered essential for impact evaluation in other disciplines (Allcott and Greenstone 2012; Hamilton et al. 2013; Vine et al. 2014; Frederiks et al. 2016), it also seems highly likely that many energy replications would fail even if attempted.

In the remainder of the paper, we focus in particular on energy efficiency research. Whilst much of what we say also holds true for any energy research, energy efficiency research is of particular importance: The International Energy Agency (IEA) labels energy efficiency as the 'first fuel' and the market for energy efficiency investments is very large – estimated between USD 310 billion and USD 360 billion in 2011 (IEA 2014). Hence, energy efficiency should be at the forefront of energy research, and it is of paramount importance it is done 'right'. The public is possibly the largest stakeholder regarding the reproducibility of science; in most disciplines, the majority of research is paid by public funds. More importantly, in applied research on energy efficiency, the public's welfare is also directly at stake. Energy efficiency research has a much greater likelihood of directly informing public policy and interventions than any lab-based study in psychology on e.g. cognitive processing of images. If we (as conductors of research that could inform government policies) installed wall insulation that would increase energy efficiency but also increase humidity and mould issues threatening health, or if we advertised wall insulation as a way to substantially reduce bills but it wouldn't deliver savings, we could harm the public beyond 'wasting' public money. Hence, we need to ensure robust research.

In the remainder of the introduction, we will lay out which factors contribute to the replication crisis and what the current status of replicability and robust science considerations is in the energy field. We then move on to describing tools that can contribute to raising the bar of energy research.

Factors contributing to the replication crisis

What are the factors that contribute to the replicability/reproducibility crisis? We argue that it is mainly due to a lack of protocols that make it easy for researchers to uphold robust research practices. Whilst this puts the integrity and quality of the individual researcher at the centre, it is not a moral judgement - all researchers are subject to institutional and peer pressure, and are as much victims and perpetrators of the system (Chambers et al. 2014). This is evidenced by the common saying 'publish or perish': The researcher who wants to maintain a successful career needs to rapidly and continually publish academic work which can come at the cost of sacrificing the quality of the work. Both research councils and journals could contribute to higher quality and replicable research, as discussed in the final section of this paper. Moreover, researchers are human beings who can and do make mistakes. The most infamous failed reproduction in economics was mostly due to a spreadsheet error (Herndon et al. 2014).

Figure 1 (from Chambers et al. 2014) shows a circle map of the scientific method and the points where questionable research practices can come into play, compromising the validity of the findings and hence increasing the threat of non-replicability. We have also introduced labels that highlight the points in the research process during which research is most vulnerable to mistakes, such as rounding errors when writing up

[2] This is a deliberately broad definition, however we recognise that there is still disagreement about whether replications should follow identical research designs, participant samples and analysis methods as the original studies or whether 'robust' results should be able to withstand minor modifications to the original procedure or analysis specification to be considered robust and replicable or whether such deviations simply reflect justifiable limits in the generalisability of results (assuming that the original results were not generalised widely).

numerical results or misinterpretation of a result due to incomplete recording of meta-data (e.g. assuming energy demand is measured in kWh per day if no time interval is specified).

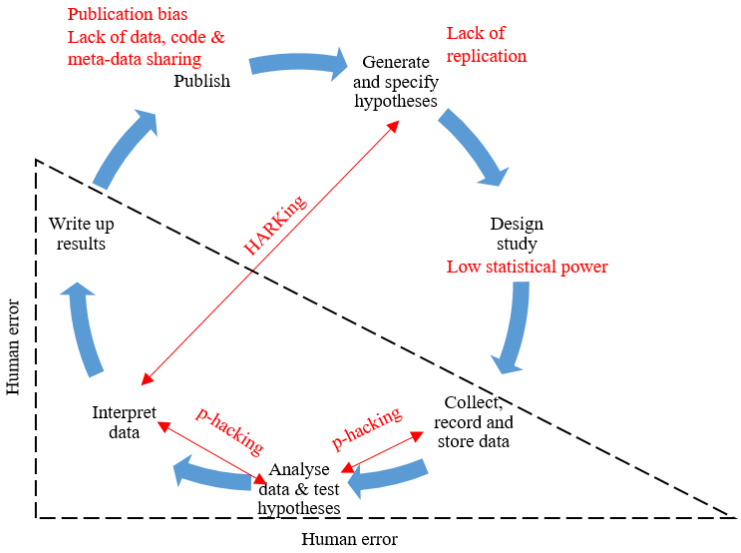

Figure 1. The hypothetico-deductive model of the scientific method is compromised by a range of questionable research practices (QRPs; red) and human error. Figure adapted from Chambers et al. (2014).

In general, there is a lack of replication which impedes the elimination of false discoveries and weakens the evidence base underpinning theory. Pure replication studies are rarely done given that they in the current research climate, they have a much lower status than novel discoveries and are much harder to publish.

Poor study design yields low statistical power. An underpowered study, i.e. having too few participants / observations, increases the likelihood of missing true findings and reduces the likelihood that a statistically significant result reflects a true effect (Button et al. 2013). P-hacking means taking liberal decisions with the data, for example, by exploiting researcher degrees of freedom. This can take various forms: stopping data collection once analyses return statistically significant effects, excluding participants whose data does not conform with the overall observed (or desired) trend, omitting some experiments or variables from the analysis, outlier manipulation[3], turning continuous into categorical variables, etc. The aim is almost always to achieve significant results which are usually indicated by a p-value < 0.05, and which increase likelihood of publication.

HARKing stands for 'hypothesizing after results are known', and means generating a hypothesis from the data but then pretending one had had it all along, turning exploratory into confirmatory data analysis and giving greater credibility to the results. However, not all replication issues arise due to deliberate manipulation. Mistakes can appear in the rush to write up results for publication; while few researchers would expect their manuscript to be free of typos unless it was proof read, perhaps very few are in the habit of asking colleagues or co-authors to check their code prior to submission.

Lack of data sharing prevents the detection of fake data. Fabricating data is less common than other practices with a meta-analysis showing that about 2% of researchers admitted to having fabricated, falsified or modified data or results at least once and up to 33.7% admitted other questionable research practices (Fanelli 2009).

The state of the art in energy research

As already alluded to above, energy research currently shows little concerned about any potential replication crisis and has few mechanisms in place that might prevent it. For example, of the 700+ subscribers to the Center for Open Science's Transparency and Openness Guidelines (Open Science Framework 2014) – a set of transparency and openness guidelines to which journals are invited to subscribe – less than 0.1% are energy or environment journals. Google scholar (and Mendeley) finds one hit when searching for "energy efficiency" and

[3] Outlier correction per se can be a useful and necessary tool, e.g. to detect erranous measures. However, the procedure for doing so should be specified in advance and not upon receipt of the data where the temptation might be to apply a criterion such as to get rid off or keep values for favourable research outcomes.

"replication" in the title (Scott 1997), a study that was published in 'Energy Economics'[4], a journal leaning more towards economic research which has higher research standards.

Despite this overall inertia to react to and be proactive about the replication crisis observed in other fields, there have been some recent publications and activities on methodological challenges in the energy field. The importance of addressing methodological challenges and creating better research has been highlighted by Sovacool in his inaugural (and highly cited) article for the first issue of 'Energy Research and Social Science': As one of the research questions to be addressed he stated "How can researchers minimize bias—their own, and that of their subjects − when doing research?'(Sovacool 2014, p. 11) – bias-free research is a prerequisite for robust and replicable research.

A recent paper argued that randomised control trials (RCT) should be used to evaluate energy efficiency policy, and that protocols or guidelines should be introduced to increase the rigour of energy efficiency evaluations (Vine et al. 2014). This view has been expanded in another recent article presenting practical guidelines for designing, conducting and evaluating the impact of energy-related behaviour change programs (Frederiks et al. 2016). The authors also emphasized that experimental research, i.e. RCTs are generally the best approach for obtaining valid estimates about the effectiveness of behaviour change programs, and that research up to date has often been below standard to draw causal inferences. Yet, we stress that performing an RCT per se, even a well-designed one, will not guarantee results that are 'true' and can be replicated – if for example, still cherry-picking for results. To avoid this, in paper at the BEHAVE conference 2016, we argued the need to specify analyses *before* carrying out the research, and registering them publicly to ensure analytical transparency (Nicolson et al. 2016).

There has also been some recent recognition of the importance of replication by a journal in the energy field. 'Energy Economics' issued a call for contributions to a Special Issue on replication in December 2016. Papers eligible for inclusion are all economics papers with some relation to energy. In addition, the journal has created a 'replication paper' as a new type of submission. Whilst the journal comes less from an energy and more from an economics perspective, which is further advanced in questions about replicability, this is nonetheless a promising step.

Whilst there are no direct replication studies, there are several areas where the same concept was tested in multiple studies, with meta-analyses or review articles synthesising the findings (e.g. Staddon et al. 2016; Abrahamse et al. 2005; Davis et al. 2013). It might give some initial consolation – if studies testing the same concept albeit operationalized and assessed differently in different samples find the same basic results, there must be some merit in the findings – however, it still depends on the quality of the individual studies if the findings are indeed sound. A recent meta-analysis of 32 North American interventions aimed at reducing residential electricity use (involving in-home displays, dynamic pricing and automated devices) indicated that almost all studies suffered from biases, the most common being the volunteer bias: in ~85% of the studies the sample consisted of volunteers (Davis et al. 2013). Bias, and how to avoid it, is discussed in more detail below.

Research methods for robust impact assessment

Many research questions in the energy efficiency area hinge on whether some new policy or intervention works – that is, questions which require us to quantify the causal impact of one independent variable (e.g. installation of external wall insulation) on another dependent variable (e.g. energy demand). The way to answer these questions is not through case studies, interviews, focus groups or machine learning algorithms[5]. Instead, we need research methods designed to measure causal effects (or causal 'impacts'). As noted in Frederiks et al. (2016), the best method for measuring causal effects is a randomised control trial. However, there are a number of different methods for estimating causal effects to which we refer as 'impact assessment methods' (Table 1).

[4] Search conducted on 9 December 2016.

[5] Whilst case studies, interviews and focus groups may help you to understand *why* the policy or intervention is or is not working (as estimated using an impact assessment method), these qualitative methods are not on their own able to identify whether a program has a large, small or no impact on your dependent variable. Whilst machine learning algorithms may be used to develop an intervention (e.g. to provide tailored feedback to households on their energy use based on smart meter data), the algorithms themselves cannot be used to determine whether the program or intervention itself will reduce energy demand.

Table 1: A summary of methods for robust impact evaluation from most to least robust.

Impact assessment method	Variations of the method	Removes or reduces omitted variable bias	Removes/reduces selection bias	Removes/reduces researcher bias or 'statistical cherry picking'
Systematic review	Rapid Evidence Assessment	n/a	n/a	No
Randomised control trial	Randomised control trial with blocking Randomised encouragement design Factorial design Step-wedged design Cluster randomised crossover design Randomised adaptive design	Removes	Removes	No
Quasi-experimental design	Matching Comparative interrupted time series Regression discontinuity Differences-in-differences Instrumental variables	Reduces	Reduces	No
Before and after design /comparison across recipients/participants versus non-participants/recipients	With control variables (e.g. multiple regression analysis) Without control variables Panel/cohort studies Time-series cross sectional Time-series Cross-sectional Case study	No	No	No

The key difference between the methods in Table 1 is the extent to which they control for selection bias and omitted variable bias. For example, imagine a government is interested in measuring the impact of a smart meter combined with an In-Home Display on energy demand in homes. In some countries, smart meters/IHDs are being rolled out to all households as business as usual unless the household rejects a smart meter. Therefore, comparing energy demand across households with and without smart meters will not give us an unbiased estimate of the effect of smart meters because people who reject a smart meter may differ systematically in their energy use to those who accept smart meters ('selection bias'). Another problem is that energy demand in households is influenced by a host of different building and occupant characteristics as well as a range of external factors such as the weather, not all of which it will be possible to measure and include in an analysis. This makes it hard to isolate the individual effect that the smart meter/IHD combination is having on energy demand, excluding all the others factors, without which the results would be vulnerable to 'omitted variable bias'. A well designed randomised control trial eliminates omitted variable bias and selection bias completely because randomisation ensures that the units (e.g. people) allocated to the treatment group and control group are statistical similar , meaning that the only difference between two or more groups in a randomised control trial is that one received an intervention and the other did not.[6] Using the example above, the causal impact of a smart meter/IHD on energy demand is the difference in average energy demand across the treatment and control groups when the experiment ends.

However, randomised control trials are not always feasible and sometimes alternative designs would, on balance, be preferable. In these circumstances, the next best alternative is to use a variation on the 'pure' randomised control trial or to use a quasi-experimental design. For example, if the government decides that it cannot randomly assign some households or regions to receive a financing scheme for household energy efficiency retrofits and others not, it may decide that it can stagger the roll out of the scheme across regions at random (a step wedge randomised control trial) or use eligibility for the finance scheme as a proxy (or 'instrument') for receiving the intervention (instrumental variables design). However, it is very important to bear in mind that, unlike randomised control trials, quasi-experimental designs require us to make additional and often very strong assumptions which may not always hold.

Given the limitations of quasi-experimental designs, we distinguish three cases in which impact evaluation by quasi-experiment should be chosen over a randomised control trial: (1) when the independent variable of interest

[6] For practical information on how to design and analyse results from a randomised control trial in the energy domain we refer readers to Frederiks et al., (2016).

is a fixed attribute that cannot be randomly assigned e.g. gender, age, dwelling type; (2) when there are high financial or practical costs involved in random assignment, such that the costs of running a full randomised control trial outweigh the benefits associated with obtaining a perfectly unbiased causal effect e.g. because there are economies of scale associated with non-random assignment; (3) when violations to the experimental procedure in a randomised control trial mean that assignment can no longer be treated as truly random[7]. [8] Although this seems like a very limited set of exceptions, there will be many times in which these exceptions are applicable. For example, just looking at (1), a major point of interest in relation to many energy efficiency interventions is the heterogeneity in impacts, either across different types of dwellings (solid wall versus cavity wall homes) and demographic sub-groups (high versus low income). As critics have pointed out (Deaton 2009; Deaton & Cartwright 2016), without modifications to the 'pure' design, randomised control trials only provide an unbiased estimate of the *average* impact of a policy or intervention. However, policymakers often need to understand the distributional implications of major policies just as much as academics want to understand the potential limits of the interventions being tested. As a result, the analysis of many randomised control trials often involve some degree of 'non-experimental' analysis. These results are therefore correlational and must then only be interpreted causally with utmost caution. If the assumptions required for a quasi-experiment do not hold, but a randomised control trial is also not feasible (see items [1]-]3] above) then before and after comparisons or comparisons across self-selected groups of participants vs non participants may provide valuable evidence that cannot be obtained by any other means. Most treatment-effect heterogeneity effects are assessed non-experimentally.

However, although the methods outlined above can minimise selection bias and omitted variable bias, none can eliminate the potentially much greater and harder to quantify biases created by conscious or unconscious biases on the part of the researchers designing the experiments, analysing the data or conducting the systematic review. As discussed in the introduction, these biases undermine the replicability of research and therefore the ability of policymakers to make sound policy decisions. The next section discusses tools that have been used in other domains (mostly clinical research) to address the issue of researcher bias and which we argue must also be adopted in energy efficiency research. Adaptations of these tools for energy research are also outlined.

Tools for increasing research replicability

Pre-analysis planning and trial registration

Two key practices undermine the replicability of research: (1) the 'file drawer' problem, whereby results that do not exceed conventional thresholds required for statistical significance (e.g. $p<0.05$) are less likely to be submitted or accepted for peer-reviewed publication; (2) 'fishing', whereby researchers (or research funders) consciously or unconsciously select analysis specifications that support their prior beliefs or desired conclusions (Lin & Green 2016). To mitigate the 'file drawer' problem, medical researchers are required to pre-register clinical trials in advance on searchable trial registries. Registering the trial in advance makes it possible for anyone to find the trial regardless of whether the study or results are subsequently published. Some trial registries prompt researchers to update the registration with the results, which is invaluable for carrying out complete systematic reviews. To minimise the 'fishing' problem, clinical researchers are required to produce analysis plans which outline what the key outcome measures are and a step-by-step outline of how the analysis will be conducted, including what covariates and statistical tests will be employed to estimate treatment effects, how missing data will be handled and how key variables will be coded etc. These 'pre-analysis plans' (PAPs) are then pre-registered with the trial and given a time-stamp and registration number.

---

[7] Randomised control trials can go wrong, for example, sometimes people in the control group may end up receiving the intervention allocated to the 'intervention group'. If planned effectively, the results could then be analysed as quasi-experimental research programmes. For example, if uptake of the intervention is non-random due to interference between treatment and control groups, the experiment could be collapsed into an instrumental variables design, in which the offer of the intervention is used as a proxy (an 'instrument').

[8] We also acknowledge that there are many cases of policies or programmes having been constructed in such a way that precludes impact evaluation as if the data were generated from a randomised control trial e.g. if people have selected themselves into the intervention group or where everyone gets or is eligible for the intervention (Vine et al. 2014). However, we strongly recommend that energy efficiency programmes and policies are designed to permit impact evaluation from the outset by the most suitable (most robust but feasible) method outlined in Table 1.

Although both these tools originate from clinical research, researchers from economics (McKenzie 2014) and political science (Lin & Green 2016) are increasingly calling for these practices to be implemented in their own disciplines, following a spate of failed replications. However, to our knowledge and as argued elsewhere (Nicolson et al. 2016), these tools are not being used in energy research.

In our experience, these tools do not require substantial modification to be relevant in energy research. We have designed a number of trials which we have pre-registered along with their respective PAPs. Although not a pre-requisite for registering a trial (we have used egap, a political science registry), having an energy specific trial registry could help to raise the profile of pre-registration and thereby increase the adoption of such practices.

One practical barrier to the adoption of PAPs in energy research is that, unlike clinical research and to a certain extent political science and economics, there is much less agreement over how key outcomes should be measured and what analysis methods should be used for identifying specific effects (Nicolson et al. 2016). However, giving thought to how variables will be coded and measured in advance is only likely to increase the quality of the subsequent research. Further, since these decisions will have to be made at some point in the research process anyway, PAPs merely redistribute this part of the research workload from the middle of the research process to the beginning – which arguably has other key advantages, such as considering potential pitfalls to planned data collection / analysis at a time when they can actually be changed (Nicolson et al. 2016).

Another barrier to the use of PAPs, but which is certainly not unique to energy research, is that it can be difficult and time consuming to make the plan robust to unexpected events such as new data, unexpectedly high levels of missing data or unexpected data distributions. However, we argue that pre-analysis plans should not be used to make it 'mandatory' for researchers to follow their PAPs unquestioningly regardless of future events. They should merely make it incumbent on the researcher to highlight what analyses were planned and justify any variations in the analysis, to minimise the risk that analysis decisions are being chosen to suit particular conclusions or theoretical standpoints. For example, imagine that a researcher planned to use OLS regression to estimate the causal effect of independent variable *x* on dependent variable *y*. Once the data is received, they note that dependent variable *y* has a strong positive skew. Most statistical textbooks will advise that, in this case, the variable should be log transformed. It would be statistically incorrect for the researcher not to modify the analysis to incorporate this 'unexpected' finding and make a note of why this was done in the journal article or any other write up of the results. However, an even more reliable approach would be for the researcher to 'fall back' on a set of Standing Operating Procedures which define how data will be treated in the event of non-normal distributions (Lin & Green 2016).

Research synthesis

Rigorous, transparent research based on the approaches described so far is fundamental to permitting effective evaluation of energy efficiency interventions. However, in all but the very largest examples, individual studies will be limited in the extent to which their results can be generalized across populations and similar interventions, and over space and time. Generalizability and robustness can be enhanced through synthesis of the findings of multiple studies in the form of a review, which can also try to account for the variation in findings that study replication can produce. However, reviews take many forms, and the means by which studies are identified for a review, and their findings synthesized, is an important determinant of the kind of conclusions that can be drawn. The following sub-section briefly sets out the principal characteristics of 'systematic' review methods, often viewed as the 'gold standard' approach to such synthesis.

Systematic reviews 'seek to collate *all* evidence that fits pre-specified eligibility criteria in order to address a specific research question … [they] aim to minimize bias by using explicit, systematic methods' (Higgins & Green 2008). Originally developed to inform evaluation of the effectiveness of pharmaceuticals and clinical practice, systematic reviews (and variants on the method) have been applied in an increasingly wide range of other subjects, including energy (e.g. Warren [2014]; Staddon et al. [2016]). The basic characteristics of systematic reviews are as follows (developed from Higgins & Green [2008]):

- *Clearly stated objectives and approach.* Systematic reviews usually start out with one or more research (review) questions and objectives. It is standard practice to prepare (and ideally publish) in advance a protocol setting out in detail in the approach that will be taken. This includes, for example, detail on search strategy and eligibility criteria for included studies. Published systematic reviews also include a detailed method section reporting the procedure as it was in fact conducted.
- *Systematic search.* A full systematic review attempts to identify all relevant evidence, whether published or unpublished, or in academic or grey literature. This is achieved by searches of databases and other electronic resources, hand searches (including of reference lists) and stakeholder engagement. Search terms, locations, times and results are recorded and reported.
- *Assessment of validity*. Based on the reported characteristics of studies, an assessment is made as to their degree of validity and relevance to the review question(s). A range of tools are available to assist this

depending on the type of research. Consideration is given to factors such as sample size, recruitment method and randomization procedures.

- *Systematic extraction, synthesis and presentation of study characteristics and findings*. Details of each study are recorded according to specified criteria, and the findings synthesized (taking account of assessed validity) to produce an overall assessment of the state of evidence relating to the review question(s). This may involve statistical meta-analysis for quantitative studies.

Compared to non-systematic reviews, following this approach means bias in selecting and reporting evidence is at least made explicit, and ideally minimized. However, as the foregoing points should make clear, full systematic reviews are resource-intensive and may take a team of reviewers years to complete. A range of other approaches can be employed which differ in various respects and in their resource requirements, and which may be better suited to meeting certain requirements. Most relevant in the context of energy efficiency are as follows (see Grant & Booth [2009] for a full discussion):

- *Rapid evidence review/assessment*. Follows the systematic review approach but adapts to fit resource constraints – usually availability of time. For example, studies may only be included from the past five years, or only in OECD countries. This approach is often seen in a policy context where quick answers are needed in specific areas of policy.
- *Systematic/evidence map*. Usually more superficial extraction of study characteristics and focus, with the primary aim of identify subjects either for full systematic review (where there is a sufficient concentration of research) or for more primary research. May be used as the first stage of a full systematic review, and to inform 'gap analysis'.
- *Systematized review*. Draws on systematic review methods to inform a standard literature review, unlikely to assure comprehensiveness.

It is also important to point out that, while they were originally developed to synthesize the findings of RCTs, systematic review approaches have been developed to accommodate a wide range of methods and epistemologies (see Gough et al. [2012]). For example, realist synthesis is used to explore how certain mechanisms work in different contexts, rather than focusing on specific interventions (Pawson 2002). As mentioned above, systematic review approaches are increasingly being employed in the context of energy demand research. However, their use is still limited compared to other research areas such as health. Work is currently underway in our team to map the use of systematic review approaches in energy research and consider the specific challenges presented by their use in this area. These are anticipated to include the diversity of disciplines and study types involved, and the possibility of transferring findings usefully between national contexts (where, for example, climate and regulatory regimes may be very different).

Reporting research

The full, precise and accurate reporting of research is a key link in the chain to facilitate not just replication, but also synthesis and interpretability of the work. This sub-section briefly expands on the reasons for this and presents examples of reporting guidelines and checklists employed in other research domains. It is self-evident that in order to permit faithful replication of a study, enough detail must be provided of the method and context of the first study that the replication can be designed appropriately. Likewise, the replication study must provide similar detail to allow useful comparison of the findings and make clear the possible reasons for differences. The ability to make useful comparisons between studies is also essential from the point of view of synthesis which, as explained above, relies on reported study characteristics to assess risk of bias and weight of evidence. Finally, even when considered in its own right, a report of research must provide sufficient information about the conduct and analysis for readers to decide whether they believe the conclusions to be well-founded.A key challenge for researchers can be deciding what constitutes 'sufficient detail' in reporting. To reduce subjectivity in such decisions, sets of guidelines have been developed (often in the form of checklists) specifying what aspects of studies should be reported. Since research approaches such as randomized trials and systematic review come from health research, this area has produced most of the leading reporting guidelines. The specific guidelines differ according to type of study (see Table 2 for details of guidelines for some main study types).

Table 2: Leading reporting guidelines by study type.

Study type	Reporting guidelines	Notes
Randomized trial	CONSORT	Requires flowchart of phases of trial and includes 25-item checklist (see (Schulz et al. 2010)
Systematic review	PRISMA	Flowchart and 27-item checklist (see (Liberati et al. 2009)
Predictive model	TRIPOD	22-item checklist (see (Moons et al. 2015). See also (Bennett & Manuel 2012) for alternative proposals.
Qualitative study (interviews and focus groups)	COREQ	32-item checklist (Tong et al. 2007)

Further research is required to determine the extent to which reports of energy efficiency research contain the information required by the guidelines listed. However, initial searching suggests that the specific guidelines themselves have little currency in energy research. Focusing on CONSORT, for example, a full-text search of the bibliographic database Scopus for references to the term "CONSORT statement" or "CONSORT 2010 statement" identified over 15,000 articles classified under 'medicine', over 600 under 'social science' (although primarily with a health focus), over 100 under 'engineering', but just two under 'energy'. This is likely due to a combination of factors. The CONSORT guidelines were developed with health research in mind so it is unsurprising that they are predominantly represented there. Journals publishing research in this area often have a requirement for the guidelines to be followed. There is a much greater preponderance of the use of randomized trials compared to in the study of energy efficiency. Nevertheless, it is striking that, given the established dominance of CONSORT as a reporting standard, there is almost no reference to it in the energy literature.

We suggest that applying the reporting rigour seen in other fields to the sharing of findings on energy efficiency is both feasible and essential. Existing statements and checklists provide a useful guide; discussion is needed to what extent these need to be adapted (or new guidelines developed) to work in the study of energy efficiency.

Publication of datasets with computer code in data repositories

Making data publicly available is a key aspect in preventing data fraud and in promoting scientific integrity and replication studies. Errors (intentional or unintentional) in analysis could be identified by others. The real strength of publication of datasets and computer code is in combination with PAPs: If neither data collection tools, variables to be used in analyses nor exclusion criteria have been pre-specified, uploaded data could still be manipulated. Together with PAPs, though, cherry-picking of results and other forms of p-hacking could be uncovered. There are further benefits to depositing data: e.g. it increases the visibility of the research as data sources can be cited, it is economic as data be re-used, it can give further discoveries that the original authors had not thought of, and it can help to establish collaboration.

The webpage http://www.re3data.org/ provides a list of more than 1,500 research data repositories that cover research data repositories from different academic disciplines; however, this includes institutional webpages which cannot be accessed by everyone. The webpage is funded by the German Research Foundation (DFG), and is mentioned by the European Commission in its document "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020"(European Commission 2016). Nature Scientific Data mandates uploading the data underlying manuscripts on submission in appropriate public data repository. They host a list of recommended data repositories on their webpage (http://www.nature.com/sdata/policies/repositories#general). Repositories included on this page have been evaluated for data access, preservation and stability. Currently (December 2016), there are no recommended repositories for energy research, but several for social science research such as Harvard Dataverse, and UK Data Service ReShare. OpenEI, Open Energy Information, is a US based repository supported by the US government where researchers can upload their data for free. Uploading data is quite straightforward, however, it needs to be ensured that confidentiality of research participants is safeguarded, i.e. all data need to be de-identified. Name, date of birth, and address would always need to be deleted or at least to be reduced in precision, e.g. reducing date of birth to year only. Anonymization is more problematic for qualitative data where e.g. video or audio recordings have been used. The UK Data Archive gives a good overview on what to consider in anonymization (http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation?index=0). Also, a meta-document or 'ReadMe' document needs to be prepared to help others to understand variables measured.

A particular challenge for energy efficiency research is that projects are often done with a commercial partner, e.g. a utility company who for competitive reasons might oppose publication of data sets. In the ideal case, this would be addressed prior to commencing any research and an agreement reached on which meta-data (if not the total data) could be shared. Another challenge is that energy efficiency studies can result in very large data sets which can incur heavy charges. Whilst many depositories are free to use, other (such as Dryad) charge in general and in particular for large data sets. Any studies involving smart meter data for a large number of homes, could end up with hundreds of gigabytes or even terabytes of data. For example, Dryad charges $50 for each 10 GB beyond the initial 20 GB. However, in the light of what research actually costs, the deposition cost is negligible.

Discussion and conclusion

Beyond the individual

In this paper we outlined tools that we believe can help energy researchers increase the transparency, replicability and ultimately the usefulness of their research. However, we recognize that we need a broader shift in the research ecosystem to improve the quality of research and aid replication studies. We briefly discuss here three areas where we see potential for this: academic publishing, funding and supporting infrastructure.

As already mentioned, if journals demanded PAPs and publication of data sources, including code, this would be a crucial step. They may also endorse reporting guidelines for certain kinds of paper, and request peer reviewers to check manuscripts against them. Journals can contribute beyond this: they could have explicit replication streams that only accept replication studies as contribution and special issues on replication, as the Energy Economics journal is currently doing. The journal 'Psychological Science' has developed a badge system where authors can earn 'badges': the Preregistered badge (for preregistering the design and analysis plan of the reported research and reporting the results as planned); the Open Materials badge (for making the components of the methods needed to reproduce the study publicly available); and the Open Data badge (for making the data needed to reproduce the reported results publicly available) (Eich 2014). Other journals could follow suit, if not making the requirement for open data and materials and preregistration mandatory. Another option is to grant conditional acceptance on methods and theory submitted prior to conducting that study, the condition being that the authors stick to their study plan, a format generally called 'Registered Reports (RR)' that was first used in the journal Cortex (Chambers 2013). Such submissions would help to overcome the publication bias that non-significant results are not being published.

Funding bodies could reserve funding explicitly for replication studies. The Netherlands Organisation for Scientific Research NWO has made three million Euros available for replication and reproduction studies of highly cited, influential findings in a pilot programme (NWO 2016). Other options would be to include successful replications, i.e. the applicant's own work was successfully replicated, as a CV entry and to have researchers list only their top five publications in their CV in grant applications, as is already done by several research councils, such as the European Research Council in its starting grant.

Finally, there may be a case for developing dedicated infrastructure to support research rigour and replicability in energy research. This could take the form of an online repository for documents such as PAPs and systematic review protocols with requirements specifically tailored to accommodate work designed for this topic. It could additionally include guidance and reporting checklists, making it a useful hub and resource to support researchers in planning their work and sharing their plans and findings. It would be necessary to consider the extent to which such a resource would be a useful addition to existing initiatives in related fields (for example the Collaboration for Environmental Evidence[9] for systematic reviews).

Limits and next steps

In closing, we want to highlight some of the challenges and limits to replicability and the deployment of the kinds of tools set out here. Our focus has primarily been on quantitative deductive research based around hypothesis testing. Qualitative research rarely has any explicit hypotheses which are statistically tested; hence, much of the afore-mentioned does not apply in their research. In particular, in participative studies where the researcher is explicitly part of the research process, one would not expect to be able to replicate findings. However, we believe there is still useful guidance to be gleaned from tools such as reporting checklists (to ensure that the way the work was conducted is as transparent as possible) and approaches to research synthesis designed to accommodate qualitative findings.

Also, we do not want to stifle academic discovery – sometimes the most interesting finding is the one that we have not pre-specified. Pre-specification works well if testing is done in a well-researched, theory-strong area but in novel or less-theory informed areas, it might be hard to lay out hypotheses beforehand. In such a case, it is key to be transparent about the explorative nature of the study, and clearly indicate which (if any) of the analyses were pre-specified and which ones not. Journals could, similar to a 'replication' stream, have an 'exploration' stream for such studies.

Finally, we have suggested ways in which the tools we present could be adapted, further investigated, and mainstreamed (for example through requirement by journals). However, we recognize that this is not a straightforward process. Within the energy research community there is likely to be a diversity of views as to the

[9] www.environmentalevidence.org/

appropriateness and need for such tools to be introduced, and if they are, by what means and in what form. We have already highlighted some possible issues and challenges that we perceive, but expect there to be many others which we may not even have considered. We also recognize that the vast array of formal and informal means by which such tools would need to find their way into the practice of energy research in order to be effective requires that there is sufficiently broad consensus on need and approach. Such consensus can only be reached through broad and inclusive engagement and debate. With this paper we want to begin the discussion.

References

Abrahamse, W. et al., 2005. A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, 25(3), pp.273–291. Available at: http://www.sciencedirect.com/science/article/pii/S027249440500054X [Accessed February 19, 2014].

Bennett, C. & Manuel, D.G., 2012. Reporting guidelines for modelling studies. *BMC Medical Research Methodology*, 12, p.168.

Button, K.S. et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5), pp.365–76.

Camerer, C.F. et al., 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), pp.1433–1436.

Chambers, C.D. et al., 2014. Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1(1), pp.4–17.

Chambers, C.D., 2013. Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49(3), pp.609–610.

Davis, A.L. et al., 2013. Setting a standard for electricity pilot studies. *Energy Policy*, 62, pp.401–409.

Deaton, A. & Cartwright, N., 2016. *Understanding and Misunderstanding Randomized Controlled Trials*, Cambridge, MA. Available at: http://www.nber.org/papers/w22595.pdf [Accessed September 26, 2016].

Deaton, A.S., 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development.

Eich, E., 2014. Business not as usual. *Psychological science*, 25(1), pp.3–6.

European Commission, 2016. *H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*, Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

Fanelli, D., 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5).

Frederiks, E.R. et al., 2016. Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers. *Energy Research & Social Science*, 22, pp.147–164.

Gough, D., Oliver, S. & Thomas, J., 2012. *An introduction to systematic reviews*, Thousand Oaks, CA: Sage.

Grant, M.J. & Booth, A., 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), pp.91–108.

Herndon, T., Ash, M. & Pollin, R., 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2), pp.257–279. Available at: http://cje.oxfordjournals.org/cgi/doi/10.1093/cje/bet075 [Accessed November 4, 2016].

Higgins, J.P.T. & Green, S., 2008. *Cochrane handbook for systematic reviews of interventions*, Wiley Online Library.

IEA, 2014. *Energy Efficiency Market Report*, Available at: www.1e.com/energycampaign/.../Server_Energy_and_Efficiency_Report_2009.pdf.

Liberati, A. et al., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339, p.b2700.

Lin, W. & Green, D.P., 2016. Standard Operating Procedures: A Safety Net for Pre-Analysis Plans. *PS: Political Science & Politics*, 49(3), pp.495–500. Available at: http://www.journals.cambridge.org/abstract_S1049096516000810 [Accessed August 30, 2016].

McKenzie, D., 2014. pre-analysis plans | Impact Evaluations. *Development Impact - The World Bank Blog*. Available at: http://blogs.worldbank.org/impactevaluations/category/tags/pre-analysis-plans [Accessed May 21, 2014].

Moons, K.G.M. et al., 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, 162(1), p.W1.

Nicolson, M., Huebner, G. & Shipworth, D., 2016. Applying behavioural economics to boost uptake to "smart" time of use tariffs without using opt-out enrolment: a pre-analysis plan for a randomised control trial. In *Behave 4th European Conference on Behaviour and Energy Efficiency*. Coimbra, Portugal, pp. 1–16.

NWO, 2016. NWO makes 3 million available for Replication Studies pilot. *NWO News & Events*, p.1. Available at: http://www.nwo.nl/en/news-and-events/news/2016/nwo-makes-3-million-available-for-replication-studies-pilot.html.

Open Science Collaboration, O.S. et al., 2015. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, 349(6251), p.aac4716.

Open Science Framework, 2014. The TOP Guidelines. *Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices By the Berkeley Initiative for Transparency in the Social Sciences, SCIENCE Magazine, and the Center for Open Science*, (1), pp.1–5. Available at: https://osf.io/ud578/\nhttps://osf.io/9f6gx/wiki/home/.

Pawson, R., 2002. Evidence-based Policy: The Promise of `Realist Synthesis'. *Evaluation*, 8(3), pp.340–358.

Schulz, K.F., Altman, D.G. & Moher, D., 2010. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, p.c332.

Scott, S., 1997. Household energy efficiency in Ireland: A replication study of ownership of energy saving items. *Energy Economics*, 19(2), pp.187–208.

Sovacool, B.K., 2014. What are we doing here? Analyzing fifteen years of energy scholarship and proposing a social science research agenda. *Energy Research & Social Science*, 1, pp.1–29. Available at: http://www.sciencedirect.com/science/article/pii/S2214629614000073 [Accessed October 16, 2014].

Staddon, S.C. et al., 2016. Intervening to change behaviour and save energy in the workplace: A systematic review of available evidence. *Energy Research & Social Science*, 17, pp.30–51.

Tong, A., Sainsbury, P. & Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), pp.349–357.

Vine, E. et al., 2014. Experimentation and the evaluation of energy efficiency programs. *Energy Efficiency*, 7(4), pp.627–640. Available at: http://link.springer.com/10.1007/s12053-013-9244-4 [Accessed May 24, 2016].

Warren, P., 2014. The use of systematic reviews to analyse demand-side management policy. *Energy Efficiency*, 7(3), pp.417–427.

Acknowledgements

GMH, MF, SE, and DS are supported by Research Councils UK (RCUK) Centre for Energy Epidemiology (EP/K011839/1). MLN, CH and HK are supported by the UK Engineering and Physical Sciences Research Council (ESPRC) grant numbers EP/L01517X/1 and EP/H009612/1. CJ is supported by OFGEM's LCN Fund through UK Power Networks.