Energy Research & Social Science 22 (2016) 147–164

Contents lists available at ScienceDirect

Energy Research & Social Science

journal homepage: www.elsevier.com/locate/erss

Review

Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers

Elisha R. Frederiks*, Karen Stenner, Elizabeth V. Hobman, Mark Fischle

CSIRO Adaptive Social and Economic Systems, Ecosciences Precinct, GPO Box 2583, Brisbane, QLD 4001, Australia[1]

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form 23 August 2016
Accepted 23 August 2016
Available online 27 September 2016

Keywords:
Research design
Behavior change
Interventions
Randomized controlled trial
Program evaluation
Consumer behavior

ABSTRACT

Governments and policymakers around the globe are becoming increasingly interested in how to effectively change the behavior of energy consumers. In the residential sector, numerous programs are attempting to shift the behavior of individuals and households in the public interest—for example toward more energy efficient practices, greater uptake of demand-side management technology, increased use of renewable energy, and better responsiveness to new tariffs (e.g., dynamic pricing), to name but a few. However, the effectiveness of such behavior change interventions is often limited, or even unknown, due to weaknesses in program design and evaluation of program impact on behavior. To help policymakers avoid such pitfalls, this paper outlines some practical guidelines for designing, conducting and, most importantly, evaluating the impact of energy-related behavior change programs and initiatives. We explain why randomized controlled trials are generally the optimal approach for obtaining scientifically valid estimates of a behavioral program's efficacy and effectiveness. In parallel, we offer specific guidelines for strengthening the validity, reliability and generalizability of empirical findings about program impact on behavior. Adopting these guidelines will help to improve program design and delivery, thereby allowing more accurate evaluation of the true cost-effectiveness, utility and mass-scalability of future energy-related behavioral interventions.

Contents

* Corresponding author.
E-mail addresses: elisha.frederiks@csiro.au (E.R. Frederiks), karen.stenner@csiro.au (K. Stenner), elizabeth.v.hobman@csiro.au (E.V. Hobman), mark.fischle@csiro.au (M. Fischle).

[1] Note: CSIRO (Commonwealth Scientific and Industrial Research Organisation).

http://dx.doi.org/10.1016/j.erss.2016.08.020

1. Introduction

Designing and delivering effective behavior change programs is critically important for policymakers, practitioners and researchers grappling with the challenge of shifting energy consumer behavior in positive ways, e.g., toward more energy efficient practices, greater uptake of renewables and energy-saving technology, more frequent use of low-emission transportation, better responsiveness to dynamic/cost-reflective electricity pricing, and higher participation in demand-side management, to name but a few. A range of strategies have been designed to encourage pro-environmental behavior in general, as well as household energy efficiency and conservation more specifically [1–5]. And over the years, countless studies, review papers and meta-analyses have been undertaken in an effort to evaluate the impact of these interventions on changing consumer behavior [2,5–10]. Although the literature suggests that certain behavioral strategies can be effective for motivating household energy efficiency and conservation,[1] in many cases conclusions are being drawn from studies that are not properly designed to test an intervention's precise *causal* impact on behavior. This can ultimately lead to situations where the benefits and return-on-investment of a behavioral intervention cannot be evaluated, making it extremely difficult – arguably impossible – to determine not only whether the intervention should be rolled out more broadly, but also whether it can be *cost-effectively* scaled to millions of consumers across the population.

To avoid such situations, we explain here that it is imperative to use a robust experimental design – in particular, a randomized controlled trial (RCT) where participants are randomly assigned to experimental groups (also known as 'intervention', 'trial' or 'treatment' groups) and control groups. Randomized experiments of this nature are generally the optimal design, and certainly the most robust approach, for any behavior change program one intends ultimately to validate [11–14]. The RCT approach offers the most scientifically robust and empirically defensible way of: (i) determining whether a cause-and-effect relationship exists between a particular intervention and its intended outcome, and (ii) assessing the validity, utility and overall cost-effectiveness of an intervention, relative to business-as-usual or alternative interventions. Certainly we recognize that non-randomized and non-experimental research designs may be preferable – indeed, entirely appropriate and necessary – for exploring other types of questions. However, when the primary aim is to evaluate the effectiveness of an intervention for changing behavior – that is, to determine its causal impact in a scientifically rigorous way – then designing the intervention as a RCT is the most appropriate approach.[2]

In the field of energy research, there is growing awareness of the value afforded by randomized experiments, with scholars such as Allcott and Mullainathan [15], Todd et al. [16] and Vine et al. [17] all sharing our view that, where at all possible, RCTs are the preferred design for testing the efficacy and effectiveness of behavioral interventions targeting household energy efficiency and conservation. Yet to date, this approach has not been widely applied in practice. In the next section, we present a critical review of the literature to identify some of the most common design flaws and methodological limitations that have featured in prior research, some of which cast serious doubt over the validity of key findings and conclusions drawn from studies of behavioral interventions in the residential energy domain. Following this critique of the literature, we then devote substantial attention to outlining a set of best practice guidelines that we hope will encourage more rigorous experimental research in this space. Our aim is to equip practitioners and policymakers with the basic tools needed to successfully design, implement and (very importantly) evaluate any energy-related intervention that aims to shift consumer behavior.

1.1. Behavior change research in the energy domain

Broadly speaking, behaviors related to residential energy conservation can be categorized into 'curtailment' behaviors, which are the routine, repetitive efforts to reduce consumption on a day-to-day basis (e.g., switching off lights; lowering thermostat settings; reducing the frequency, intensity and duration of appliance usage) and 'efficiency' behaviors, which are one-time actions such as purchasing new energy efficient technology and building modifications (e.g., upgrading old, inefficient appliances; replacing incandescent bulbs with new LEDs; installing fixtures like low-flow shower heads, insulation and solar panels) [18,19].[3] Over the years, various strategies have been designed to modify such behaviors. Interventions targeting household consumption and conservation have ranged from so-called 'antecedent strategies' that occur before the target behavior (e.g., providing information and education; using prompts and reminders; goal-setting and commitment techniques; and using social norms, peer influence and social modeling) through to 'consequence' strategies that occur after the target behavior (e.g., encouraging self-monitoring of one's behavior or performance; delivery of feedback; and use of rewards/penalties) (for a review, see Ref. [5]). Community-wide programs and policies that may shift consumer behavior in a desired direction are also common, including the introduction of dynamic electricity tariffs (e.g., cost-reflective pricing) providing consumers with a 'price signal' that may incentivize reductions in peak energy use; rollout of demand-side management initiatives such as direct load control; and offering other financial inducements such as rebates, subsidies, tax credits or bonuses for certain energy-saving actions. Yet to date, the methods used to design, implement, and evaluate the impacts of these sorts of behavioral strategies have not always been appro-

[1] Note, however, that evidence for the effectiveness of different interventions to promote pro-environmental behavior such as energy conservation has been far from consistent across studies. Research suggests that each behavioral strategy has boundary conditions under which it is maximally effective, such that no single technique will work for all people, at all times, and in all situations. Rather, the impact on behaviour often depends on myriad factors such as the context, the target population, the behavior of interest, among other moderating factors (for an overview, see Ref. [1]).

[2] If the aim is to evaluate *and explain* program impact, there is much to be said for integrating data from multiple sources and methods. For example, combining quantitative evidence derived from an RCT with data from non-experimental and qualitative methodologies can allow the investigator to isolate with great precision the true impact of an intervention on behavior, while also gaining deep insights into the processes underlying those effects.

[3] Note that although the energy-saving potential of efficiency behaviors tends to be greater than curtailment behaviors, the former do not necessarily lead to reductions in overall energy consumption. There is ample evidence in the literature of so-called 'rebound effects', where energy efficient measures and technological improvements may actually result in greater overall energy usage [20–23]. For example, buying more efficient appliances may not reduce consumption if a consumer then uses those appliances more often.

priate, which raises questions about the reliability and validity of results reported in some studies.

Indeed, it is clear from our search of the peer-reviewed literature that many of the behavior change studies conducted over the past few decades have not featured scientifically rigorous experiments that yield robust, unbiased estimates of actual behavior change. Historically, the evaluation of energy efficiency programs and behavioral interventions has typically involved qualitative research, observational studies and quasi-experimental methods. While there are some exceptions (for recent examples, see Ref. [24–28]), true experiments in the form of RCTs have been few and far between. Our review of the literature suggests that many intervention studies have been limited by non-randomized or non-experimental designs (e.g., relying instead on self-report surveys, qualitative interviews, focus groups, small case studies, or naturalistic observations), small sample sizes, the absence of proper control groups, confounded treatments, short timeframes (e.g., no long-term monitoring or follow-up) and/or highly selective (even self-selected) participants—all of which constrain one's ability to draw generalizable conclusions with confidence. As noted earlier, while non-experimental methods can be useful for exploring the rich detail and context of complex behavior, they are unable to determine the precise *causal* impact of an intervention in changing a particular behavior. In order to evaluate behavior change – and quantify the actual cost-effectiveness and return-on-investment of an energy-related intervention – a robust experimental design is required.

Alongside an abundance of individual studies, it appears that prior literature reviews and meta-analyses of behavior change interventions in the residential energy domain have also featured some studies with inadequate designs and methodological limitations [5,6,10]. For instance, as we illustrate below, some reviews have included studies with very small or biased samples,[4] or studies that have assessed behavior change by relying on subjective self-report data rather than objective measures of actual consumption.[5] In some instances, the specific nature and/or content of the interventions themselves are unclear or not defined. And there are many cases where multiple interventions are combined together (rather than having distinct interventions or aspects of interventions separated out for proper testing). This raises concerns over confounding of effects, and can render it difficult (if not impossible) to isolate which distinct programs, or aspects of a program, are ultimately consequential in shifting the behavior in question. Methodological issues of this nature were noted over two decades ago by Dwyer et al. [7], who conducted a critical review of behavioral interventions to preserve the environment, including studies targeting residential energy conservation. Based on an evaluation of 54 environmental behavior change studies published during the 1980s, the authors concluded that much of this early research was not designed to allow meaningful comparisons of multiple interventions, and few studies monitored longer-term behavior change. Despite some exceptions, well-designed randomized experiments were rare, and very little research was conducted in such a way as to allow for a direct, unconfounded comparison of distinct behavioral interventions.

While recent years have seen an increased focus on robust experimentation, there are still many cases where behavioral interventions cannot be evaluated in a scientifically valid way due to fundamental limitations in research design and methodology. Support for this notion can be found in a number of reviews and meta-analyses conducted over the past decade, three of which we cite here as illustrative examples. The first is Abrahamse et al.'s [5] widely cited review of intervention studies aimed at household energy conservation, which included 38 studies that were '*mostly field experiments, using quasi-experimental designs*' (p. 274). Based on our understanding of each study's design and method, less than half were RCTs. Some studies had so-called 'control' groups that were not randomly assigned, while others had no control group at all (e.g., simple pre-test/post-test designs). Of those that were randomized experiments, many had small and/or selective samples, raising questions over the reliability and generalizability of results. The authors note that alongside large within-group variances (in energy use), very small sample sizes may have reduced the statistical power of some studies. They also draw attention to other critical problems with prior research, such as cases where intervention 'effects' are reported based only on changes in self-reported behavior, as well as limited monitoring of long-term behavior change.

More recently, a comprehensive meta-analysis of information-based energy conservation experiments by Delmas et al. [6] provides further evidence of the methodological challenges prevalent in this domain. Evidence from 156 published field trials (with over half a million participants in total) conducted from 1975 to 2012 was reviewed, yet many of these studies suffered from problems such as small samples, short time-periods, no control groups, failure to take baseline measurements, low levels of granularity in outcome measures, failure to control for the impacts of potential confounding variables (e.g., weather, household demographics) and other confounding issues caused by failure to adequately separate out distinct interventions (or aspects thereof) in order to properly isolate their individual effects. Of critical importance, the authors found that the average treatment effect of information strategies on energy savings diminishes with the methodological rigor of the study – that is, a markedly lower savings effect (1.99%) was found for higher quality studies with adequate controls (weather, demographics, control group) compared to lower quality studies without such controls (9.57%). It suggests that the savings effects reported in less rigorous studies may be considerable overestimates, casting doubt on the reliability of some reported effect sizes and indicating continuing methodological issues in the current literature. Viewed more broadly, these findings are of great concern given that very important and expensive decisions regarding program utility, cost-effectiveness and scalability may be made on the basis of the 'evidence' derived from studies of this nature.

Third and finally, a recent meta-analysis by Davis et al. [10] also illustrates the limitations of prior behavior change research in this space. Their review of 32 North American interventions designed to reduce residential electricity use (involving in-home displays, dynamic pricing and automated devices) concluded that most studies were at high risk of bias from multiple sources, with the most common methodological problems including volunteer selection bias (almost 85% of all studies used volunteers, e.g., as with opt-in designs), intervention selection bias (~63% involved either the participant or researcher choosing the intervention group, rather than using random assignment) and attrition bias (~44% had data exclusions or withdrawals, and/or data not 'missing at random'). By applying the 'risk-of-bias' approach developed for medical clinical trials, Davis et al. [10] found most studies had inadequate designs and methodological features that were expected to inflate estimates of intervention effectiveness. After the authors re-calculated these estimates to adjust for this risk of bias, the resulting values were often less than half those originally reported in the studies reviewed. Similar to Delmas et al.'s [6] meta-analysis, which highlighted the potential for over-estimated treatment effects, these

[4] This reliance on small and unrepresentative samples is undesirable as it can threaten the external validity of empirical findings and lead to reduced statistical power for detecting statistically significant effects.

[5] Self-report data may be vulnerable to various cognitive biases (e.g., impression management and social desirability effects) and memory errors, which means that it may not reflect real-world behavior. If the focus of an intervention is on changing a behavioral outcome (e.g., a household's level of energy consumption), it is important to measure actual behavior (e.g., energy consumption in kWh).

results are sobering when one considers that research of this nature is used by policymakers to make important decisions about future energy-saving initiatives, including choosing between alternative investments.

Clearly then, while there is a growing body of literature on behavioral interventions to promote household energy efficiency and conservation, it appears that conclusions are often being drawn from studies with non-experimental designs, small and/or selective samples, and methodological weaknesses—characteristics that prohibit a robust scientific test of causality. Put simply, it seems that a good proportion of behavioral programs are designed and delivered in ways that constrain their ability to detect real behavior change, limiting the policy implications that can be drawn from the experience, and thus the prospect of broader societal gains. As our review of the current literature has shown, this inability to evaluate program impact may be due to one or more deficiencies in program design. The most critical of these is deploying a non-experimental approach (and/or not including a randomized control group), ultimately leaving us unable to isolate the true causal impact of an intervention. But they also include a broader set of methodological limitations such as very small/selective samples, confounded interventions, short timeframes, or using inappropriate measures (e.g., self-report data) to conclude that something has 'worked' without having objective behavioral evidence to substantiate such a claim. Unfortunately, all of this can lead to situations where critical public policy decisions are made (say, in favor of the mass roll-out of an expensive energy efficiency campaign) based on unsound or misleading evidence, or in some cases no evidence at all. In contrast (as we will argue in the remainder of this paper), designing and implementing behavior change programs that allow precise scientific evaluation of true program impact makes it relatively simple and straightforward to determine critical outcomes of interest, such as the cost-effectiveness, practical utility and scalability of the program in question.

1.2. The need for best practice guidelines for evaluating energy behavior change

In an effort to help energy policymakers avoid the critical methodological limitations we have described here, we provide an overview of key principles for the successful design, delivery, analysis and evaluation of behavior change interventions, broadly conceived. While these principles are generic and multidisciplinary in nature, and thus applicable to any intervention aimed at changing behavior, we specifically focus on the domain of residential energy use (certainly an area of public policy where more evidence-based behavioral research is warranted) in order to highlight the practical benefits and potential applications for all those seeking to positively influence the energy-related behavior of individual consumers and households. Our principles are informed by a multidisciplinary review of the behavior change literature in general, alongside the broader literature on quantitative research methods, experimental design and robust statistical analysis. In parallel, and importantly, our guidelines stem from a critical review of literature from the past four decades that includes a wealth of intervention studies targeting household energy consumption and conservation, alongside recognition of the methodological problems that have limited the capacity of many studies to empirically determine the actual *impact* of various energy-focused interventions on consumer behavior.

In devising our guidelines, we have noted the early calls of scholars such as Dwyer et al. [7] for more robust, systematic research in this space, as well as more recent reviews and meta-analyses that have flagged design flaws and methodological issues with prior intervention studies targeting household energy use [5,6,10]. Further, we aim to extend on the recent work of Vine et al. [17], who discussed the value of experimentation for evaluating energy efficiency programs. After presenting a strong rationale for using robust experiments to answer serious policy questions, and then examining the various barriers that can constrain the use of experimental designs in practice, Vine et al. [17] suggested that '*it may be prudent to develop protocols or guidelines for conducting experiments, especially to help those without any experience in experimental design*' (p. 635). We seek to present such guidelines here, specifically written for energy policymakers not abreast of the relevant academic literature, who might see the considerable benefits of conducting randomized experiments in the residential energy space.

In parallel, we also aim to address an important gap in the literature recently identified by Sovacool [29]. In the inaugural issue of this journal, Sovacool proposed a variety of methodological and topical areas for future research, including a comprehensive social science research agenda comprised of specific questions worthy of further exploration. Our paper examines issues directly related to several of the questions proposed by Sovacool [29] with a view to deepening and broadening energy research, such as: '*How can the benefits of "human-centered" research methods be best coupled with quantitative forms of data collection and analysis?*' and, '*How can researchers minimize bias—their own, and that of their subjects – when doing research?*' (p. 11). We are thus making a concerted effort to engage with what this journal considers to be the most pressing scientific questions in the energy domain. By providing guidelines that allow a broad range of energy practitioners and policymakers to properly exploit rigorous experimental designs and methods to investigate consumer behavior, we hope to encourage and enable more research into the 'human' dimensions of energy issues. These aspects are greatly under-studied relative to the more traditional focus on technological dimensions, and far less well understood (see Ref. [30] for a discussion of the importance of integrating social science into energy studies).

In outlining our recommendations, we explicitly acknowledge that there may be constraints to using randomized experiments in real-world settings. As discussed at length by Vine et al. [17], and still further in the following, there are a wide range of regulatory, institutional and design barriers, and even ethical issues (e.g., equity concerns), that can limit the use of experimental designs in evaluating the impact of energy-related programs and initiatives. We certainly recognize that it is sometimes simply inappropriate or unfeasible to use RCTs to evaluate the extent and nature of behavior change. In such cases one may have no better option than an alternative research design, such as a quasi-experiment.[6] We also recognize that the suitability of a particular research design and methodology ultimately depends on the specific question one aims to answer. As such, we do not dispute the value offered by quasi-experimental, observational or qualitative studies in many cases. Indeed, non-randomized and non-experimental designs may be preferable, and entirely appropriate and necessary, for exploring certain types of (non-evaluative) research questions. For example, if the aim of a study is to describe individual experiences, explore subjective states (e.g., feelings, attitudes, perceptions), or better understand underlying processes, it may be entirely appropriate to forego a highly controlled experiment in favor of methods that allow for richer texture and insights, greater flexibility, iterative incorporation of feedback and learning, even the researcher's active involvement in the behavior being studied (via 'participant observation'). Such research can also be considered a useful first step to inform the initial development of behavioral interventions that are then subjected to iterative testing and refinement via randomized controlled trials. Different methodologies have their

[6] Quasi-experiments are similar to RCTs, but lack the critical feature of *random assignment* of participants to experimental and control groups.

unique strengths and weaknesses, and there is no one-size-fits-all approach. Thus, while our paper outlines a number of principles for designing and delivering energy-related behavioral interventions in ways that permit scientific testing of causal impact, we understand that there will inevitably be situations where some of these principles are unfeasible or inappropriate.

1.3. Aims and objectives

With these issues in mind, and with explicit recognition of the real-world constraints that apply to RCTs, we aim now to provide a set of systematic recommendations for designing, delivering and evaluating behavior change interventions, in those cases where true experiments are indeed feasible, scientifically appropriate and ethically sound. In the sections that follow, evidence-based recommendations are proposed for all those phases of the research process that may influence a program's apparent impact, with a particular emphasis on designing the process in such a way as to preserve one's ability to rigorously determine the precise magnitude and nature of the intervention's impact in terms of changing real-world behavior. These phases include initial scoping and hypothesis formulation, experimental design, measures and methodology, data collection and analysis, and replication of findings. Each of these stages will be discussed, in turn, in order to provide a comprehensive roadmap for designing effective behavior change programs, by which we mean: programs that are driven by clear aims and objectives; that tightly align with a specific set of relevant research questions; and that produce valid, reliable and readily generalizable evidence on whether the program has been successful in achieving those objectives. We provide specific recommendations that policymakers can follow to design, implement and evaluate behavior change programs, including the important particulars around how to test whether an intervention has actually produced the desired outcome for the population of interest.

The remainder of this paper proceeds as follows. First, we discuss the importance of initial hypothesis formulation as the basis for defining a program's intended outcomes in terms of shifting some target behavior. Second, we outline several important principles of program design, with a primary focus on describing the key features of, and critical benefits afforded by RCTs. Next, we outline some general guidelines for each of the major stages of program delivery, ranging from participant sampling, recruitment and allocation, to the use of valid, reliable and standardized measures, through to collecting objective behavioral data over a sufficient time frame. We then discuss the importance of appropriate statistical analyses in determining the program's precise impact on behavior. Finally, we conclude by explaining the value of replicating experimental results, particularly when seeking to generalize the findings of a single study more broadly, or when experimental results are to serve as the basis for making fundamental changes in public policy. We propose that by following these guidelines (to the extent possible), energy researchers, practitioners and policymakers will be better placed to determine – with a high degree of accuracy – whether a particular strategy, program or intervention is a cost-effective and readily scalable means of changing behavior right across the target population. Put simply, by deploying a scientifically rigorous behavior change program, one will have greater capacity to determine which specific intervention(s) prove effective, and when, where, why and for whom this is generally the case.

A summary of our key recommendations for designing, conducting and evaluating behavior change programs is presented in Table 1. These recommendations stem from a multidisciplinary review of the behavior change literature in general, alongside well-established principles for conducting robust quantitative research and analysis, particularly in regard to randomized experiments. As such, there is no single source from which any recommendation derives, as an extensive literature underpins each one. In the sections that follow, we cite specific sources to support key concepts where relevant. However, for more in-depth guidance on each of the stages involved in conducting an RCT, see Refs. [12,31,32]. To help demonstrate key points, the final column of Table 1 provides illustrations of what one might do at each step. We continue to utilize throughout the same example of a behavioral intervention motivating households to reduce energy consumption by means of 'messaging' that conveys descriptive social norms (i.e., that presents them with information about their household's energy use for a given time period, relative to a group of similar neighbors). For real-world applications of similar sorts of interventions that have actually been tested in the field using RCTs, see Refs. [15,24,25,33], as well as our discussion of real-world randomized controlled field experiments in Section 4.

2. Key recommendations for program design and methodology

2.1. Formulation of hypotheses: clearly specify the program's objectives and predicted impact on behavior

When it comes to designing, implementing and evaluating the efficacy of a behavior change program, the first step is clearly specifying its intended effect, that is, the predicted impact of the intervention on participants' behavior. The purpose of the program should be stated in specific, measurable and time-bound terms, making explicit what criteria will be used to evaluate its success. As part of this process, key research questions and associated hypotheses should be identified (for an overview of hypothesis testing in scientific research, see Refs. [34,35]). It should be clear what outcomes one expects to have observed by the close of the program, and what new insights one aims to have learned. Hypotheses should consist of clear, concise descriptions of the predicted effect of the program (i.e., independent variable) on one or more indicators of behavior change (i.e., dependent variable(s)) and should identify the nature, direction and magnitude of expected effects. Even for completely new programs, we recommend drawing on the best available theory and extant empirical evidence to formulate at least tentative expectations about the likely impact of the intervention on one or more outcome measures.

2.2. Program design: use a randomized controlled trial to test program impact

As outlined earlier, a randomized controlled trial is a rigorous, scientific experiment purposely designed to test the efficacy of an intervention on a sample of participants drawn from some target population [12,31,32]. Although RCTs may be more resource intensive in certain situations, and sometimes difficult to implement faithfully, there is widespread consensus across diverse academic disciplines (including medicine, healthcare, social welfare, education, employment, psychology, behavioral economics and environmental sciences) that RCTs are the most valid and reliable method of evaluating the *impact* of a behavior change intervention—that is, for scientifically testing both the efficacy (the beneficial effect of a given program under ideal and optimal conditions of delivery, such as a tightly controlled trial) and effectiveness (the effect of a program under more real-world conditions or in 'natural' settings) of an intervention [36]. RCTs are frequently described as setting a methodological 'gold standard' for evaluating interventions [11], by ensuring the scientific validity of empirical findings and claims of causal inference. Over the years, a large body of research has attested to the methodological superiority of RCTs for

Table 1
Summary of key recommendations for designing, conducting and evaluating behavior change programs.

Recommendation	Key points	Example
1.Formulate hypotheses: clearly specify the program's objectives and predicted impact on behavior	• Clearly define the program's purpose in specific, measurable, time-bound terms • Explicitly define what criteria will be used to evaluate the program's success • Identify key research questions and hypotheses • Draw on the best available theory and evidence to formulate hypotheses	• For an intervention that involves a utility providing consumers with 'descriptive normative information' on their utility bills (e.g., graph illustrating their household's energy use in kWh relative to that of similar neighbors), one might hypothesize that over the life of the study (e.g., 'x' billing cycles), households that receive bills with normative information will consume significantly less energy ($p < .05$) than households that receive bills without the normative information; that is, there will be a statistically significant difference in the mean energy use of households that do vs. do not receive bills with normative information, with the former households consuming significantly less (on average) than the latter.
2.Program design: use a randomized controlled trial to test the program's impact	• Randomly assign a representative sample of participants to alternative intervention ('treatment') and control groups, to ensure the different groups are 'equal on average' to begin with. Do not allow participants to 'self-select' one group or another (assignment must be purely random). Ensure there is at least one randomly assigned 'control' group that is not exposed to any intervention • If the intervention, alternatively, is 'opt-in' then all participants who wish to take part in the intervention should be assigned by a purely random process to either the experimental or control group. That is, it should not be the case that those who failed to consent/respond to the offer, or otherwise indicated they did not wish to participate in the intervention, are considered the 'control' group. Rather, it must be the case that of those households that wish to participate, some are randomly assigned instead to the control group, and receive no 'intervention' as such	• From an entire population of residential energy consumers in the utility's service area, households are randomly assigned to receive either an energy bill that includes descriptive normative information (i.e., the experimental or 'treatment' group) or a standard, business-as-usual bill that does not include this information (i.e., the control group)
3.Methodology: ensure sound participant recruitment, construct measurement and data collection	• Determine how sample representativeness (degree to which participants in sample accurately reflect the relevant 'population') will be defined • Determine the minimum sample size needed to ensure sufficient statistical power and accuracy of estimates • Draw a sufficiently large and representative sample of participants from the target population, e.g., use proportional sampling methods (random sampling) that reduce bias and sampling error • Ensure intervention fidelity and standardized delivery to participants. Where practical and appropriate, avoid exposing participants to multiple treatments, either simultaneously or sequentially • Use valid, reliable measures that are appropriate for gauging behavior change over time to collect objective behavioral data on key outcome variable(s) • Collect baseline (pre-intervention) data on key outcomes of interest for all participants. For energy usage studies, ideally at least 12 continuous months of historical data (i.e., one year of energy use data, prior to intervention) should be collected, to ensure baseline consumption reflects seasonal effects • Ensure a length of data collection sufficient to assess the durability and persistence of any observed effects. That is, monitor long-term behavioral outcomes to test whether short-term effects remain stable or change (perhaps diminish) over time	• Determine the population of interest for the study. For example, it might be all residential energy customers for a particular utility, or all energy consumers in a particular geographical region, or all customers on a particular (e.g., dynamic) electricity tariff. Determining the appropriate population from which to draw one's representative sample will ultimately depend upon the nature of the research question being investigated • Wherever possible, energy consumption data for all households (treatment and control groups) is collected prior to intervention, e.g., historical data from one or more years before any normative information is presented • Over the life of the intervention, the energy utility continues to unobtrusively monitor and record the energy use of all households in exactly the same way, e.g., with the same frequency and duration of data collection, and using identical methods and measures. To examine the durability of the intervention's effect over the trial period, household energy consumption is monitored repeatedly over time using direct and objective measures, e.g., electricity/gas meter readings • At the end of the intervention when the provision of normative information is discontinued, the monitoring of household energy use continues unobtrusively for several more billing cycles in order to test whether the intervention's effects on household consumption persist after the delivery of normative information ceases

4.Data analyses: conduct appropriate statistical analysis of program effects	• Validate that treatment and control groups are equivalent across key baseline characteristics of interest (e.g., pre-intervention energy use, geographic locations, household and dwelling features), particularly in terms of variables that may affect a participant's response to the intervention • Conduct preliminary analyses to review simple descriptive statistics, and evaluate the proportion and patterns of missing data. Rule out systematic patterns in missing data that might otherwise introduce bias into the analyses, confirming that data are 'missing at random'. Otherwise take immediate steps to rectify data deficiencies to the extent possible, or else consciously address/make appropriate allowance for those known deficiencies in the analyses • Determine the size and significance of any differences between treatment and control conditions, for example by conducting *t*-tests of differences in group means, or estimating the unstandardized regression coefficient attaching to membership of treatment vs. control group • Tests of significance should always be accompanied by expressions of effect size, which is usually the figure of key interest and practical utility for subsequent policy formulation and selection among alternative interventions or investments • Report results in a suitable unit of measurement of central interest and utility to the relevant audience (e.g., actual change in kWh usage per day/month/quarter), ensuring that the key outcome variable is a valid and reliable, objective and readily observable measure of the actual behavior of interest • Simple before-and-after comparisons of the outcomes of a single intervention group cannot determine causal impact, due to the wide range of alternative explanatory variables to which any observed change in behavior could be attributed (in the absence of a control group baseline), e.g., simple 'maturation' over time, media coverage, political/economic events occurring in the interim, unseasonal weather. Any observed change in the behavior of those exposed to an intervention must always be compared to that evident in a randomly assigned (and therefore equivalent/'equal-on-average') control group	• Prior to analysis, an equivalency check should be conducted to validate that the control and intervention groups are comparable across key baseline characteristics. Assuming a sufficiently large sample is drawn, random assignment is deployed correctly, and there are no systematic patterns subsequently evident in missing data, the treatment and control groups should start out 'equal-on-average' at the outset • Endeavour to ensure that non-response/missing data is not systematically more likely among (say) high usage/energy-inefficient households that in the course of the trial/intervention may have suffered 'high bill shock' and (in consequence) abandoned the utility and therefore dropped out of the study, i.e., by switching energy retailers. This would constitute data *not* 'missing at random' (particularly to the extent this was more likely to occur in a particular experimental group) and pose serious risk of introducing bias into the analyses. This might then provide a misleading picture of the true impact on energy consumption of (say) providing households with normative messages about their energy consumption relative to their neighbors • At the end of the intervention period, data analysis should be undertaken (including formal tests of statistical significance and determination of effect size) to compare pre- and post-intervention consumption data for households in the experimental and control groups, to identify the effect on subsequent energy consumption of the descriptive normative information
5.Replication: Retest/repeat the program to gain stronger evidence of its impact	• Repeat the trial, including repeatedly measuring the same outcome variables – first, in the same environment and context, and (ideally) then in different environments and contexts – to identify any sources of variation and to isolate the true treatment effect with greater precision and confidence • Exploit all available opportunities to replicate results using different samples of participants, at different points in times, across different settings, and using different measures and methods	• Assess the generalizability (external validity) of the findings, e.g., determine whether descriptive normative information has the same effect for households in different geographic regions, climate zones and socio-cultural environments, for customers of other energy utilities, and/or at different points in time. Repeat the normative messaging intervention in other settings and with entirely different samples of residential energy consumers

testing program efficacy and effectiveness, particularly in, but not limited to, the medical and healthcare domains [13,14,37–39]. As Solomon and colleagues [12; p. 6] described it:

> 'RCTs are considered to be the "gold standard" of evidence when determining the effectiveness of policy and practice interventions... Intervention science currently revolves around the RCT as *the* standard of evidence... Randomized trials establish evidence for particular populations and practice contexts. Although other research strategies may have some capacity to build an evidence base for an intervention, RCTs provide the strongest supporting evidence universally recognized to establish effectiveness.'

The central feature and core methodological strength of RCTs is random assignment of participants (after initial recruitment and assessment of eligibility) to alternative treatment and control groups, which allows confident claims to be made about causal relationships, i.e., about the true impact of a behavior change intervention on a specific outcome of interest [40,41]. Random assignment helps to ensure that the different groups are 'equal on average' – rather than systematically different – across all other variables[7] (whether observed or not) having potential to influence the outcomes of interest, apart from the intervention itself, whose impact can thus be 'separated out'[8]; [12,32,42]. By virtue of random assignment, any measured difference in the dependent variable(s), known as the average treatment effect, can be attributed to the intervention, as the groups on average do not start out varying on any other variable. This key attribute of randomized experiments is known as 'internal validity' [34,43]. Random assignment of participants to treatment and control conditions neatly eliminates rival explanations for any differences subsequently observed between treatment and control groups, allowing us to make confident, and very precise claims about the efficacy of an intervention.

The core logic of randomized experiments is simply this: if a treatment is assigned at random, assignment of that treatment cannot be systematically related to any other variable, and its effects on the outcome of interest can thus be separated out from the effects of all other variables (whether known or unknown, observed or unobserved) with which it might otherwise have co-varied. With the treatment and control groups 'equal on average' at the outset, the control group effectively controls for the changes that would have occurred in the absence of intervention, since human behavior change is continuous and has many causes, including prior motivations and intentions; learning and adaptation; social, economic and political events; seasonal and other changes in both the natural and built environment; natural maturation and the simple passage of time.

Contrast the foregoing with non-randomized research designs in which people can – whether it is the investigator's intention or not – 'self-select' (rather than being randomly assigned to) one condition or another, such that (for example) especially motivated, or knowledgeable, or 'needy' people end up seeking out a particular treatment, and especially disinterested and unwilling participants (or worse still, those who refuse to consent to participate) end up being assigned to the so-called 'control' condition (for more detail on selection bias, see Refs. [44–47]). Without random assignment, these treatment and control groups would be systematically different to begin with, prior to intervention. Consequently, it would be impossible to separate out the effects of the intervention (on any outcomes subsequently observed) from the influence of those participant characteristics that were systematically over-represented in one or another condition at the outset. The effect of the intervention itself would be 'confounded' with other rival explanatory factors, such as participants' motivation, knowledge, needs, interests, willingness, *and* everything that might conceivably co-vary with those attributes. Thus, using randomization to assign treatments to participants is critical for reducing bias [48,49]. For some real-world examples of selection biases in energy research, see Davis et al. [10] who have described cases of volunteer selection bias (where those who volunteer for a study differ from the broader population) and intervention selection bias (where participants can choose their preferred treatment group, rather than being randomly assigned). In earlier work, Hartman [50] has also discussed self-selection bias in the evaluation of voluntary energy conservation programs, noting that this may lead to upwardly-biased estimates of program effectiveness if participants and non-participants differ in observed demographic and economic characteristics.

In the end, non-randomized methodologies – including simple before-and-after (pre-post) comparisons, cohort studies and quasi-experiments, as well as naturalistic observations, case studies and surveys – cannot definitively rule out the possibility that observed effects are due (wholly or in part) to extraneous variables unrelated to the intervention, which are 'confounded' (tangled up and confused with) the effects of the intervention itself [41,48]. One must bear in mind that correlation does not indicate causation [51,52]. In simple survey research, for example, one might observe that householders who attended a community energy-saving workshop (e.g., a public training course to learn 'tips' to save energy) were more likely subsequently to reduce energy consumption, without ever being able to determine the precise extent to which that effect was due to experiencing the workshop itself, as opposed to being the kind of person (e.g., highly self-motivated) in the kind of circumstances (e.g., supportive family/friends) that made one more likely *both* to seek out and complete the workshop, *and* successfully reduce consumption. To enable a precise causal claim, one would need to randomly assign householders to either experience the energy-saving workshop *or not*, so as to increase the likelihood that householders in both groups are, on average, equally motivated and supported. In short, one would need a randomized controlled trial.

All that said, as we explicitly acknowledged in the introduction, sometimes it is simply not feasible, ethical or even possible to deploy a RCT, and in such cases, it may be necessary to forgo a true experimental design and instead use alternatives such as quasi-experiments. For instance, compared to individual- and household-level interventions, community-level interventions such as mass-marketing campaigns or public programs already available are less well-suited to evaluation via pure RCT designs [53].[9] Threats to external validity (generalizability of results to the

[7] Such as psychological attributes and predispositions, socio-demographics, and other key characteristics that may be associated with the outcome in question, including prior behavior and exposure to other relevant experiences.

[8] If this optimal approach of random assignment of participants is simply not possible, a minimally acceptable alternative is to appropriately 'match' participants in treatment and control groups across key characteristics, especially those factors that may be related to the outcomes of interest, such as baseline measures of key outcomes (e.g., energy consumption), prior experiences (e.g., of equivalent, related or alternative treatment), socio-demographic factors (e.g., age, gender, socio-economic status, employment status, occupation, education, household type/size, family structure), internal psychological factors (e.g., knowledge, attitudes, intentions), situational factors (e.g., geographic location, climate, built and natural environment) and other contextual factors (e.g., socio-cultural, economic, political, legal, institutional influences).

[9] To experimentally evaluate the effect of a community-wide mass-market campaign, for instance, one would need to sample and randomly assign a sufficiently large number of 'communities' to treatment and control groups, then ensure that any intervention effects do not 'spill over' from communities in the intervention group to those in the control group – something which could occur if the communities were within close geographical proximity, or if the intervention cannot be confined

broader population of interest) may also arise if the experimental setting is so tightly controlled and artificial that it no longer accurately reflects the real-world environment. Thus in many cases, researchers and policymakers face the practical challenge of striking the right balance between what is ideal in theory, and what is achievable in practice. Vine et al. [17] provide a comprehensive review of the real-world constraints that can apply to attempts to deploy experimental designs for energy efficiency programs, covering everything from regulatory and institutional barriers through to design and scope/theory barriers. We do not provide a detailed examination of those obstacles, to avoid duplicating an excellent discussion already available in the literature. But we do recognize that myriad constraints operate in the real-world. There will inevitably be times when the guidelines outlined herein, while ideal in principle, are difficult to follow in practice. For more specific recommendations about how to address some of the key barriers to conducting RCTs in applied settings, see Vine et al. [17].

2.3. Methodology and measures: ensure sound participant recruitment, construct measurement and data collection

In order to fully capitalize on these benefits of an RCT design, the next important steps are to strive for scientifically sound (ideally, random) sampling of participants, standardized delivery of the intervention, valid and reliable measurement of key constructs, and appropriate data collection. Our 'best practice' recommendations for each of these aspects are as follows.

2.3.1. Sample and recruit a sufficiently large and representative set of participants from the target population

Using an RCT design is an important first step to ensuring that the impact of a behavior change program can be properly assessed. It is also critical that appropriate techniques are used to sample and recruit participants from the target population. A fundamental prerequisite for proper sampling – but one that is often overlooked – is clearly defining the 'target' population. One needs to determine first exactly *whose* behavior is of interest, which will often also necessitate carefully specifying the location and/or relevant timeframe. Clearly defining the population of interest is inextricably linked with the program objectives and research questions one is seeking to answer, and helps ensure that the subsequent processes of participant sampling, recruitment and allocation are properly and effectively conducted.

Once the target population is clearly defined, sample size and sample representativeness are of paramount concern [54,55]. The number of participants selected and allocated to treatment and control groups should not only be large enough to ensure adequate statistical 'power' for detecting the hypothesized effects, but also sufficiently representative of the target population (as well as any smaller sub-populations where segmentation is deemed important). Generally speaking, larger samples will improve the precision of empirical results and increase the likelihood that (true) intervention effects will be detectable in the data (for more detail on sample size determination, see Refs. [56–59]). A number of factors should be considered to determine sample size requirements, such as the variability (in key attributes and outcomes) within the target population from which the sample is drawn, and the desired level of accuracy and certainty in the estimates. With inputs such as these, one can calculate the number of people required in the end sample and the number to be initially approached, and achieve the level of accuracy and certainty actually required for one's particular purposes.[10]

There is no simple rule-of-thumb for determining the required sample size, as each study is unique and a number of factors must be considered when calculating the number of participants needed to ensure sufficient power for statistical analyses and reliable estimates. So one study may only require a few hundred participants to yield adequate levels of statistical precision and ensure that the true effect of an intervention on the target population can be detected (sufficient statistical 'power'). But the minimum sample size required for another study may run into the thousands, given a different target population, objectives, design, interventions, methods or critical outcome measures. Generally speaking, larger samples will be required for longitudinal studies where participant non-response and attrition are likely to be high, where outcomes are highly variable, or where an intervention is expected to have only modest effects and/or to impact only (or especially) a small or difficult to reach sub-group of the population.

In all cases, prior to analysis one should examine the degree to which the sample drawn is representative of the target population across key variables and characteristics of interest. In the case of energy-related behavioral interventions, this may include not only objective measures of energy consumption and/or demand, but also variables that can affect consumption/demand, such as socio-demographics, household or family characteristics, dwelling/building features, geographic location, weather and climate, and any other variables that might conceivably impact a participant's response to the intervention.

That said, even if the initial and final samples are sufficiently large, threats to both external *and* internal validity may still arise if there is selective attrition of participants from comparison groups – that is, if the loss of participants across groups is systematic rather than random – as when participants prove more likely to drop out of one group (e.g., the control group) than others, and/or the participants that drop out of one group differ systematically from those that drop out of another [42]. In the case of dynamic electricity pricing experiments, for example, it is not difficult to imagine a situation where consumers who are randomly assigned to one type of tariff (e.g., a pricing scheme with high penalties for peak energy use; with no rebates or discounts for reductions in demand; with highly variable prices) are more likely to withdraw because they are receiving no benefit (or even incurring costs/penalties). Those assigned to another type of tariff (e.g., a pricing scheme with bonuses, lower penalties, and more predictable pricing) might be more likely to remain in the study. Attrition bias may be reduced by offering equal extrinsic incentives to both treatment and control groups, such as financial reimbursement and other material rewards for ongoing participation (e.g., completion bonuses). But it may still be difficult to prevent systematic drop-out of participants across groups of a magnitude sufficient to introduce significant biases into the investigation.

In any case, it remains important to strike just the right balance between sample size and sample representativeness. While the latter is not always reported in behavior intervention studies [60], neither should be neglected. One might be tempted to pursue a larger sample size by focusing in on a subset of the target population that is easier to reach—for example, those who have already indicated some interest in the topic at hand by belonging to a relevant organization (e.g., not-for-profit bodies that advocate for environmental conservation and energy efficiency); subscribing to

to a single area (i.e., whole-of-market advertising campaigns) or media market. In such cases, one may have no choice but to deploy another research design.

[10] One might also consider the aims of the research, the relative importance of the variables of interest, the likely non-response, the sampling methods used, and any practical issues or resource constraints (e.g., time, money, personnel) that apply to the study.

a related newsletter, mailing-list or blog; or being part of an online community or social network with a vested interest in the topic—as with non-probability sampling techniques such as convenience and accidental sampling. Although it depends in part on exactly what one is trying to achieve with the intervention, increasing sample size at the expense of sample representativeness is rarely a good trade-off. The sample needs to be as representative of the target population as possible if one wishes to generalize the study findings to the broader population in similar contexts, which is usually the case.[11] The extent to which the results of a particular experiment hold across variations in time, samples, settings, and so forth is known as 'external validity' [42,43,62–64]. Generally, the best way to achieve a representative sample is by using a probability sampling technique [65], where each member of the population has a known, or calculable, non-zero probability of selection. While various types of probability sampling exist, one of the simplest methods is to draw participants at random by some entirely chance mechanism from a large sampling frame that adequately captures the target population (known as 'simple random sampling'), such that each member of the sampling frame has an equal chance of being selected for the trial.[12] Random sampling guards against the introduction of bias in who is included in the study, as each 'unit' in the population (e.g., each energy-consuming household within a specified region) has an equal and known probability of being selected for participation.

Of course, not all potential participants who are initially sampled will agree to be recruited into the trial. Thus, recruitment processes can also reduce the size and representativeness of the final sample, potentially threatening the external validity of empirical findings. So in designing the recruitment channels that will be used to attract prospective participants to a behavior change program, these too should endeavor to draw in as representative a subset of the target population as possible, rather than have disproportionate appeal to certain segments therein – for example, those who are more computer literate, as in the case of online-only recruitment techniques (for reviews of strategies to improve recruitment into RCTs, see Refs. [66–68]). Again, what is at stake there is the external validity of any observed effects of the intervention, should one wish to generalize from the study findings to the prospects for a mass roll-out. To this end, one should also carefully consider the vehicle(s) through which participants are likely to be recruited in any real-world roll-out of the program, and perhaps endeavor to replicate in the trial those recruitment channels that would actually be deployed, particularly when targeting difficult to reach populations. In that case, it will be critically important to randomly assign alternative treatment(s) and control conditions *within* each channel. Otherwise, one risks confounding the impact of the intervention itself (upon the outcomes of interest) with the impact of the channel by which certain participants were recruited, e.g., confusing it with the impact of socio-demographic attributes prevalent among participants recruited by a particular route.

If multiple recruitment channels (e.g., face-to-face, online, telephone, mail-out) are used to attract prospective participants, it is important for the process to be as similar as possible across these channels, especially if (in practice) it proves unavoidable that those recruited via one mode rather than another are more or less likely to be assigned to a particular treatment or control group, i.e., if group allocation is associated with one's recruitment channel, rather than randomly assigned. For example, in the case of a residential energy conservation program, if those recruited online are more likely to be assigned to receive energy efficiency messages via digital channels than those recruited via another means such as mail-out, then the particular recruitment channel – and all that this is associated with in terms of the characteristics and circumstances of those participants – is confounded with a certain treatment, making it impossible to cleanly separate out the effects of the intervention itself.[13] As noted above, such problems are easily avoided if respondents *within* each channel can be randomly allocated across treatment(s) and control conditions. Anything short of this represents a conflation of recruitment channels with treatments, and introduces a confound that limits the confidence with which one can draw causal conclusions, although we acknowledge that in practice this is sometimes difficult to avoid. In such cases, it will be critical to *statistically* control for participants' varying recruitment channels in any analysis that subsequently attempts to isolate the impact of the alternative intervention(s) themselves.[14]

2.3.2. Ensure intervention fidelity and standardized delivery to participants

Ensuring that a behavior change intervention is delivered as intended is essential to preserving an investigator's capacity to compare equivalent, replicable interventions across time, contexts and populations [69]. The degree to which an intervention's core components have been implemented (and differentiated from the control condition/s) as originally planned is known as 'intervention fidelity' or 'treatment fidelity' [70–72]. To improve intervention fidelity and reduce the introduction of extraneous 'noise' that can serve to mask the effects of an intervention, standardizing each condition is imperative—i.e., the design and delivery of each behavior change treatment should be the same for *all* participants who are assigned to it (e.g., all participants in an energy efficiency workshop should receive the same instructions and educational materials, delivered in the same way). That is, each treatment should be highly consistent across time, place, facilitator, and so forth. Any marked deviation from standardized design and delivery has the potential to introduce additional unexplained (and likely, unexplainable) variation, because the experience of participants assigned to the same condition would essentially vary [69,70]. While the presence of random variation does not invalidate a study, by introducing 'noise' it can undermine one's ability to detect significant intervention effects (i.e., it affects measurement reliability and statistical power). Efforts should therefore be made to systematically evaluate intervention fidelity and, where possible, include fidelity data in the analysis of an intervention's outcomes.

Distinct concerns also arise, this time more a threat to internal validity, wherever participants have the ability to self-select or

[11] If one wants to generalize *beyond* the target population and to other real-world contexts, replication of experimental results across diverse samples and different settings, at varying time-points, and using a range of methods and measures is necessary [61].

[12] Note, however, that other random sampling methods such as stratified sampling may be preferable over simple random sampling in certain situations. For instance, if one is specifically interested in examining the impact of an intervention on particular sub-groups within the population, and there is a chance that simple random sampling might, by chance, miss one or more of these groups, it might be preferable to divide the population into groups (strata) first and then draw samples from within these strata.

[13] This concern dissipates somewhat if the recruitment channel is actually conceived as part of the intervention itself, e.g., if an energy utility was comparing the effectiveness of a fully digital (end-to-end) energy efficiency campaign to that of a traditional mail-out of energy efficiency information along with customer billing (but see cautions to follow in Section 2.3.3 regarding multiple interventions).

[14] Thus, if for some reason this conflation cannot possibly be avoided, a fallback (although clearly inferior) position is to carefully record the channels by which different participants were recruited. In a second-best world, this at least allows such influences to be statistically controlled for (e.g., by conducting separate analyses for participants recruited via different channels, or simply including the channel as a moderating variable in interaction with other explanatory factors). Good record-keeping will also allow uptake rates via different recruitment channels to be accurately calculated, which at a minimum may provide useful learning for the roll-out of future programs.

modify certain aspects of a treatment to meet their individual needs or preferences—for example, if participants can choose the number, type or schedule of treatments they receive, or the specific setting and environmental conditions under which they receive it.[15] Here again, the experience of an intervention is not the same across all participants who are assigned to a particular condition. But in this instance – and far more problematically – that varying experience is not random but, rather, systematically related to certain participant characteristics (e.g., to knowledge, motivation, interests, personality, geographical location, etc.). This then constitutes a threat to internal validity. This situation is generally a far more serious matter, since it has the potential to actually mislead us about program efficacy, by confounding the influence of pre-existing participant characteristics with the impact of the program itself.

2.3.3. *Where practical and appropriate, avoid exposing participants to multiple treatments, either simultaneously or sequentially*

Ideally, to ensure a clean test of an intervention's impact, it is important to deliver each treatment in isolation to a group of participants, rather than in combination with other treatments (e.g., mixing different components together). Unless a behavior change program is purposely designed to have several components working in unison, multiple treatments should not be delivered to a single group of participants as this risks 'multiple-treatment interference', whereby the effects of one treatment interact with the effects of another [73–75]. Exposing participants to more than one treatment – either simultaneously or sequentially – makes it impossible to separate out and identify the precise effect of each treatment (for examples of this issue in the domain of pro-environmental behavior experiments, including energy conservation interventions, see Refs. [5,8,76]). For example, if a behavior change program comprising four distinct aspects (e.g., education materials, goal setting, feedback, and rewards) is found to have a positive effect on participant behavior, this effect could only be broadly attributed to *something* among the mix, i.e., to the program overall. It would be extremely difficult, if not impossible, to determine whether one aspect of the program (e.g., information) was more effective than another (e.g., feedback), or whether they had additive, interactive, or even counteractive effects. For example, it could be that most of the behavior change observed was due to just one aspect, which would suggest that the others could be excluded from the program without compromising its impact, thereby enhancing cost-effectiveness. It could even be that one (or more) of the treatments has a negative influence, perhaps inhibits the effects of the other treatments, such that the program's overall impact could be even greater by removing those treatments with deleterious effects.

Thus, to accurately identify the impact of a single treatment (or to compare the relative effectiveness of several different treatments), it is necessary to carefully separate the different components that may be influencing behavior, and randomly assign them independent of one another. If it is important to also examine the interactive (vs. additive) effects of combining different aspects—for example, to determine whether the impact of treatment 'A' is enhanced (or diminished) when combined with treatment 'B', rather than administered in isolation—then a 'crossed' or factorial design should be employed. Here, the individual treatments, as well as combinations of treatments, would be randomly assigned to different groups of participants (e.g., one group might receive treatment 'A', one group receives treatment 'B', and one group receives both 'A' and 'B', while a control group receives neither).

Sometimes it is simply impractical, or otherwise inappropriate to separate out and compare the independent effects of highly interrelated and interdependent treatments—for example, goal-setting activities are frequently paired with self-monitoring, while feedback strategies are often accompanied by rewards [8]. If one has little choice but to assess the combined effect of multiple behavior change strategies mixed together, it is still necessary to keep in mind the above caveats, i.e., that the simultaneous or sequential delivery of multiple treatments to the same group of participants will place constraints on how data are analyzed, how results are interpreted, and what conclusions can be drawn about the behavioral impact of a multi-treatment program. It will not be possible to compare the relative effectiveness of the different parts of the program in isolation, and consequently, to determine the most cost-effective part (i.e., the most impactful strategy from the broader program of multiple strategies) for shifting the behavior of interest.

2.3.4. *Collect objective behavioral data using valid and reliable measures of behavioral outcomes*

To accurately assess the success of a behavior change intervention, it is also important to gauge its impact on actual *behavior*—that is, on one or more observed behavioral outcomes that are objectively measured, ideally in a real-world setting. Participants' self-reported perceptions of their own (or others') behavior (e.g., their subjective appraisals of their post-program performance) are vulnerable to well-known response biases and distortions [77–81]. This may be especially true where participants know, or think they know, what is the 'right' (i.e., socially desirable) thing to say or (claim that they) do, by virtue of completing a program whose goals or intended outcomes were manifest throughout. Such biases are even more likely when professing the desired behavior will earn them status or rewards, be they material (e.g., monetary bonus or gift) or otherwise (e.g., praise or recognition). Self-reported behavior may also be influenced by features of a data collection tool or measure, such as question order and wording, response format (e.g., open versus closed questions; rating scale design), and question context [82].

To circumvent these risks to accurate program evaluation, it is important (wherever feasible) to collect objective outcome data using valid and reliable measures of actual behaviors, and ideally including multiple measures of different aspects of those behaviors (e.g., measures of the nature, frequency, intensity and/or duration of some real-world behaviors in the relevant domain) so as to examine consistency in results and allow better measurement of the underlying constructs of interest [69,83]. Reliability describes the degree of consistency and stability in measurement; that is, the reproducibility of a measurement when it is repeated under the same conditions [34,84]. A reliable measure is one that measures the variable of interest precisely and consistently. An unreliable measure could be so imprecise or inconsistent that the 'noise' in the measure drowns out the 'signal', such that the expected effects of an intervention are less likely to be detected.

Yet generally speaking, accurate program evaluation is threatened more by invalid than unreliable measures, since failing to detect an intervention's effect is generally less problematic than being misled to believe an ineffective intervention has had an impact. A valid measure is one that actually measures what it intended, and nothing besides; that is, the extent to which the actual observed measurement directly assesses the variable of interest [69,85]. For example, in the case of a residential energy efficiency program, a valid outcome measure might be one that gauges a participant's actual electricity consumption in kilowatt hours (as indicated by third-party metering) rather than reflecting (additionally, or instead) something else entirely—like whether the

[15] Some related concerns will be raised later in the article in regard to the importance of 'intention to treat' analyses.

participant is the kind of person who responds helpfully to an investigator's request for follow-up data, is sufficiently well-organized to locate their latest electricity bill, knowledgeable enough to decipher it, and/or honest enough to report it back faithfully, even if it reflects poorly on their post-program performance in achieving the desired objectives. Validity refers to the 'truth' (vs. precision) of measurement – the extent to which one is actually measuring what one intended to measure – which in this case is participants' post-program electricity consumption (rather than their helpfulness, organization, knowledge, or honesty). To gain some perspective on the hazards of an invalid measure, consider how easy it would be to (erroneously) conclude that an energy efficiency program had discouraged electricity consumption, when all it had truly discouraged was honesty: reducing some participants' willingness to accurately report back to the investigator outcomes that they had come to understand were socially undesirable.

One should strive, wherever feasible, to collect direct and objective measures of the relevant real-world behaviors, which are standardized across all participants (including the control group). These should be gauged with minimal reliance on participants' self-reports or observations (and ideally with minimal or no awareness of being studied, although this does raise some very difficult ethical issues that we will return to later). In our energy efficiency example, this would be metered electricity consumption data for both treatment and control participants, recorded by a third-party such as the utility. Importantly, if the central aim of an intervention is to change household energy use, it is vital to measure *actual* consumption as opposed to other behaviors that can only be presumed to relate to consumption (e.g., purchasing energy efficient devices, turning off appliances, adjusting thermostats), because behavior change can occur without resultant reductions in energy use. For example, a person may upgrade to new energy efficient appliances but then increase the intensity, frequency and duration of using these appliances, such that their level of consumption actually increases. Indeed, there is extensive evidence in the literature of 'rebound' effects, where energy efficiency measures and technological innovations that reduce energy service costs, can actually lead to greater overall consumption [20–23].

In terms of measurement methods, data should be collected by a third-party – ideally one that is 'blind' to the research hypotheses, or at least, wholly disinterested – and in a minimally intrusive way (e.g., via business-as-usual processes), to ensure that the observed behavior accurately reflects how the participant would ordinarily behave when not under observation [86–88].[16] Observational methods (e.g., surveys, interviews, focus groups) that are more interactive or intrusive have the potential to inadvertently influence behavior during and after the study. Simply knowing that one is under observation or taking part in a research study may influence behavior due to so-called 'reactivity' and 'demand characteristics'—that is, experimental artifacts where participants' reactions are impacted by their mere awareness of the study (e.g., its intended aims, hypotheses, procedures, anticipated findings, etc.) and what (they assume) this implies for how participants should ideally behave [43,87–90]. For instance, participants may be susceptible to impression management and social desirability effects, and strive to play the role of a 'good subject' to satisfy the perceived needs of the researchers (see also the 'Hawthorne effect', which is a type of reactivity where participants change/improve their behavior due to awareness of being observed [91–95]).

Perhaps most problematically, if participant observation and recording is far more obtrusive in the treatment than the control group, one will have introduced a 'procedural confound' that makes it impossible to separate out the effects of the treatment itself from the impact of knowing that one is being observed [43]. Ultimately this can threaten both internal and external validity, as we will know neither the precise impact of the intervention itself on the outcomes of interest, nor the extent to which we can expect those effects to generalize to a mass roll-out, once program participants are no longer being intensely and obtrusively observed in a study.

Certainly, if a program aims or claims to shift real world behavior such as actual household energy use (e.g., consumption in kWh), then it is insufficient to only collect data on participants' self-reported actions (e.g., subjective appraisal of one's own behavior or performance), knowledge (e.g., awareness or understanding of a problem), cognitions (e.g., attitudes, intentions) or affective states (e.g., emotions).[17] As one cannot guarantee with confidence that these will lead in the end to real and consequential behavioral change, they provide insufficient and incomplete measures of program success if the ultimate goal is to change human *behavior* in a way that leads to reductions in energy consumption. Across a range of domains, research shows that there is often a discrepancy between people's knowledge, values, attitudes and/or intentions, and their subsequent actions, i.e., there may be a knowledge-behavior gap [96–99], a value- or attitude-behavior gap [100–103] and/or an intentions-behavior gap [104–106]. There is general consensus in the literature that attitudes and intentions are often poor and only indirect predictors of behavior. For example, a meta-analysis by Sheeran [105] estimated that intentions explain only about 28% of the variance in future behavior, with several groups of variables (e.g., behavior type, intention type, properties of intentions, and cognitive and personality variables) moderating the intention-behavior relationship. And a subsequent meta-analysis to examine the overall impact of changing behavioral intentions on subsequent behavior change found that a medium-to-large sized change in intention leads to only a small-to-medium change in behavior [107].

Notwithstanding the foregoing cautions, there will inevitably be situations where it is simply not feasible to capture the real-world behavior of interest, in which case one would be well advised to revert to what Aronson and colleagues [61] have called 'behavioroid' measures (certainly in preference to retrospective self-reports of behavior). 'Behavioroid' measures are designed to reflect "subjects' *commitment* to perform a behavior, without actually making them carry it out" [61][61; p. 271]. For example, rather than trying to measure energy conservation by having participants indicate their agreement with survey items that purportedly reflect their values in regard to saving energy, or relying upon them to accurately report their past performance of energy conservation actions, one might instead experimentally engineer a realistic situation where subjects must actually display their intentions or willingness to engage in such actions by 'enrolling' in, say, an energy efficiency program or demand-side management scheme such as direct-load control. One might ask consumers to (for example) sign something that indicates their '"expression of interest" in partic-

[16] We do realize that ethical (and practical) constraints may prohibit researchers from monitoring the electricity usage of households without their knowledge and informed consent. However, in our experience a waiver of consent can often be obtained from institutional ethics committees if measuring the key outcome involves only regular billing data that would be collected by the retailer in the normal course of conducting business, and that billing data is not merged for analysis with any personal information (including attitudinal data) that may have been collected without explicit consent.

[17] Self-report measures such as those typically collected in surveys, interviews and focus groups might still be used in a program evaluation to complement or support more objective outcome measures. Even then, one should always endeavor to frame and collect these self-reports (e.g., of participants' attitudes, intentions or behavior) in real-time, i.e., to ask participants to report how they are thinking, feeling and behaving *at that time*, rather than retrospectively (or prospectively). As we have noted, self-reports are prone to response biases and distortions, which include primacy and recency effects, in addition to social desirability bias.

ipating, or to subscribe to a mailing list of those wanting to be notified when the scheme in question becomes available. While we recognize that signing up for (say) an energy efficiency program is not the same as actually participating in the program, sometimes practical or ethical constraints may make it necessary to employ second-best alternatives of this nature to estimate the program's *likely* impact on consumer behavior.

In any case, it is imperative that the measures used to assess all variables – both explanatory and response/outcome variables – are demonstrably valid and reliable for the specific population of interest; that is, they should measure the variables that they purport to measure, and in a precise and accurate way. This not only applies to direct and objective measures of actual behavior, but also any indirect and subjective measures. For instance, if a survey or questionnaire is used to collect data from a specific sub-group of the population, its item and scale scores should demonstrate sound 'psychometric' or measurement properties across various indicators of reliability and validity (e.g., internal consistency, test-retest reliability, content validity, convergent and discriminant construct validity) for the sample of participants to whom it is administered [86,108]. Note that constructing new measures 'from scratch' is far from a quick and easy feat. Rather, it is a complex, time-consuming and potentially error-prone process (particularly when abstract variables are involved) that requires a high level of technical skill and expertise (for guidance on survey and scale development, see Refs. [109,110]). Where possible, rather than developing new measures oneself (i.e., using self-generated questions that have not been pre-tested), we advise practitioners to refer instead to published research to find suitable, standardized instruments whose validity and reliability as measures of the relevant constructs have already been rigorously established [111].

When the aim is to evaluate the impact of a behavior change program (and especially where the cost-effectiveness of any future mass roll-out is a central concern), we strongly recommend wherever feasible that real-world behavioral outcomes of the program are measured using direct, objective and well-validated quantitative indicators, rather than relying on indirect, subjective, idiosyncratic and qualitative ones.

2.3.5. Ensure an appropriate duration of data collection

In designing behavior change programs, investigators should ensure that they retain the ability to gauge not just the immediate impact of an intervention, but also the ongoing impact as the intervention is continued over time (i.e., durability) and the impact after the intervention is discontinued (i.e., persistence).[18] Realistic calculations of cost-effectiveness and mass scalability may often require follow-up data to be collected on participants for some time after their exposure to the intervention. In regard to exactly how long and how often one ought to monitor the behavior of participants, it depends on striking a careful balance between the accuracy of impact assessment and, again, the risk of altering the (real or apparent) impact of an intervention by the obtrusive act of monitoring it. The latter may draw participants' attention to the actions that are desired/expected of them ('demand characteristics') and thereby increase participants' motivation to perform those actions, thus muddying the effects of the intervention itself. Again, this may be particularly true if the intrusiveness of monitoring is greater for treatment than control participants, as is often the case in follow-up data collection. Whatever long-term 'tracking' of behavior is undertaken, it is important that data are collected as unobtrusively as possible [87,88] and, certainly, in exactly the same manner, and with the same frequency and duration, for treatment and control group participants alike.

Arriving at the right balance will ultimately depend on such things as the type of behavior being monitored, the nature of the participants, and the kind of intervention being tested. A lengthier period of data collection is particularly important if the behavioral outcomes in question tends to fluctuate over time (e.g., where the behavior is subject to seasonal influences) and/or is prone to a 'rebound effect' (where the behavior may revert back to pre-program levels), such that initially observed effects can essentially diminish or disappear over time. If these effects are substantially weakened or even extinguished over the observation period, this may inevitably impact assessments of the program's utility and cost-effectiveness. Generally speaking, it is also advantageous to collect historical energy consumption data for the period before an intervention begins, so as to establish a baseline measure of household energy use. To ensure baseline data reflects seasonal effects, at least one full year of retrospective data (i.e., 12 continuous months immediately prior to intervention) is recommended [16].

2.4. Data analysis: conduct appropriate statistical analyses of program effects

If a behavior change program has been well designed at the outset, then analysis of its impact on behavior is generally straightforward. It is often mistakenly assumed that statistical analysis is the hard part of program evaluation, when in fact no statistical 'bag of tricks' can salvage any usable lessons from a poorly designed intervention in which the effects of rival explanatory variables (including pre-existing participant characteristics) are confounded with the impact of the intervention itself. But if random assignment to treatment and control conditions has been successfully accomplished to begin with, then it is typically a simple matter to detect the effects of those different treatments (relative to the control group) and make confident claims about the program's impact on the outcomes of interest. In this way, RCTs have strong internal validity. In short, if one randomly assigns participants to X, or not-X, and those in the former group end up with significantly different levels of Y than the latter group, then this can be attributed to their experience of X since, by virtue of random assignment, everything else about the two groups was 'equal on average' at the outset.[19]

An important first-step to any analysis is undertaking a so-called 'equivalency check' to ensure that the treatment and control groups were comparable at baseline (i.e., prior to intervention) across key variables and characteristics of interest—for instance, overall energy consumption and billing outcomes, load profiles, socio-demographic attributes, household characteristics, dwelling features, geographical location and any other factors that are expected to influence response to the intervention being tested. Drawing a sufficiently large sample and strictly adhering to a RCT design should help to ensure that groups are equivalent at the outset; nevertheless, it is always important to formally validate.

Assuming treatment and control groups are equivalent at baseline, detecting the size and significance of any differences between experimental conditions can then be as simple as conducting *t*-tests of differences in group means, or (what amounts to the same thing) estimating the unstandardized ('metric') regression coefficient attached to having experienced a particular treatment rather than the control condition [86]. Unstandardized regression coefficients tend to be especially useful, delivering very directly what a program evaluator typically wants to know, which is the effect

[18] Allcott and Rogers [112] conceptually distinguished between two different temporal effects of an intervention: 'durability', which is the dynamics of an intervention's effect(s) as the intervention is sustained over time, and 'persistence', which is the extent to which these effects continue after the intervention is discontinued.

[19] Modeling rebound effects for program interventions over time can introduce extra complexity [113].

size[20] or *magnitude* of program impact, i.e., the change in behavior expected to come about as a result of participating in the program [86,114,115]. In any case, while the forms of analysis will ultimately be dictated by the particular program design and objectives, we note generally that tests of significance should be accompanied by expressions of effect size, with the latter results clearly interpreted in terms of the units of practical interest, e.g., the kilowatt hours of electricity consumption that would be saved by participating in a household energy efficiency program. Calculating the effect size in this way allows for straightforward computation of cost-effectiveness (at least where the program has known/calculable running costs), which is critical to determining the applied value and scalability of an intervention [116].

While we have emphasized this same point in various ways throughout, it is sufficiently important that it is worth reiterating here again. When it comes to drawing causal inferences, it is sometimes incorrectly assumed that a simple before-and-after comparison of outcomes for the treatment group(s) alone will suffice, without comparing these to a proper (randomly assigned) control group. Any over-time change that is observed among those exposed to a particular treatment is often simply attributed to the treatment itself, the assumption being that the treatment caused any observed behavior change. However, such assumptions are flawed. Simply observing pre- to post-intervention changes in a treatment group (and/or comparisons with some historical 'control group' presumed to be equivalent, but observed at a different point in time) is not sufficient for determining causality. In the absence of a proper control group, before-and-after comparisons might reflect *how* something has changed over time for the treatment group, but not precisely *why* this change has come to pass (i.e., to what it can be attributed) [48,117]. The 'how' refers to whether any change has occurred, whereas 'why' concerns the extent to which the program, intervention or treatment was the cause. The latter is the critical question we need to answer. In the case of an energy conservation program, for example, any observed reduction in energy consumption over time can only count as an effect attributable to the intervention to which 'treated' participants were exposed to the extent that it exceeds whatever reduction was observed in some randomly assigned control group with which they were 'equal on average' at the outset, prior to intervention.

To reiterate, simple before-and-after analyses of changes observed in a treatment group (without comparison to control group outcomes) will not partial out changes that are unrelated to the program and instead due to such things as seasonal variation, learning, maturation/aging, the simple passage of time and the unfolding of social, political and economic events. But with random assignment, since the effects of all such rival explanatory variables are operating equally on both treatment and control groups, they are effectively separated out in any treatment versus control group comparison of outcomes. This neatly isolates the impact of the program itself from any other (unforeseen or unaccounted) influences that may also be operating on the behavior of interest. In the absence of a randomized control group, it is near impossible to accurately separate out program impact from how participants might have behaved had they *not* been exposed to the program. In the presence of a proper control group, however, these analytic challenges can be avoided [12,31].

Note that the strictest[21] test of program impact will compare outcomes observed among all those originally assigned to the treatment and control groups, irrespective of whether they continued the program, completed their assigned treatment (or accepted their lack of treatment), complied with its protocols, and furnished all outcome measures. This is known as 'intention to treat' analysis [13,54,118,119]. Short of this, it may be that one has detected *not* the impact of being exposed to the program per se, but rather the impact of fully participating in and complying with the program (a less common occurrence in many situations). Since full participation and compliance are even harder to achieve in a mass roll-out than in a smaller scale trial or pilot test, anything other than a rigorous 'intention to treat' analysis risks over-estimating the true impact of the program in broader application, and can limit the generalizability of findings.

2.5. Replicate findings: repeat the program and re-evaluate impact

While a well-designed and sufficiently powered RCT is critical for evaluating the effectiveness of a behavior change program, an important final step – and one that is often overlooked by researchers and practitioners – is replication: the repetition of an experiment to gauge consistency of findings. Even if the experimental setting is highly similar to real life, and the study is relatively large and broad, it is still only a single study, subject to variation and error, being conducted with one sample of participants, at one point in time, and in one context. Until an iterative program of systematic replication is undertaken, some questions will inevitably remain over the extent of generalizability to other contexts [61]. Replicating an experimental study and measuring the same outcome variable(s) repeatedly helps to identify sources of variation, more precisely estimate the true treatment effect(s), and ultimately strengthen the reliability and external validity of experimental findings [42]. This is particularly important when one aims to generalize the results of a single study more broadly, and when experimental results serve as the basis for making fundamental changes in public policy. Researchers should always strive to replicate experimental results, ideally using different samples of participants, at different points in times, across different settings and situations, and using various measures and methods. If similar results are found across these variations, this offers strong evidence that a particular treatment or intervention is truly effective.

3. Caveats

While the overarching premise of our paper is that RCTs are critical for evaluating the impact of behavior change interventions, we readily acknowledge that there are inevitably situations where it is impractical and/or unethical to adhere strictly to this classic experimental design. In some cases, RCTs may be best practice for achieving scientific objectives, but not best practice from an ethical, practical, legal, social or political perspective (for a review of ethical issues in the design and conduct of RCTs, see [120]). And as noted earlier, there may be a range of regulatory, institutional, design and scope/theory barriers to using a true experimental design [17]. Arguably one of the most important issues is the fairness and ethicality of withholding effective interventions from certain participants, in pursuit of scientific rigor. This is particularly critical in the medical and healthcare domains, where one must consider whether there is truly the 'clinical equipoise' – genuine uncertainty over whether a treatment will be beneficial, or superior to

[20] The *size* or magnitude of the relationship between X and Y is related to, but distinct from the *strength* or certainty or 'tightness' of the relationship (the latter reflected by correlation coefficients, for example). The strength of a relationship reflects how *certainly* the behavioral outcomes will follow upon program participation, more than how *large* the behavioral impacts are expected to be. It is generally (though not always) the case that in program evaluation, we are interested more in the size than the strength of a relationship.

[21] But see our caveats (to follow) regarding ethical concerns around 'intention to treat' analyses.

alternative treatments – necessary to provide an ethical basis for assigning participants to different treatment and control groups. If sufficient evidence accumulates of a treatment's efficacy, there are strong ethical arguments for breaking protocol at that point and administering the treatment to all participants, at variance with a strict RCT design. While this may reduce the certainty with which one can assert causal claims and capacity to determine long-term effects of the treatment in question, it is nevertheless imperative to always adhere to best practice standards for the ethical conduct of experimental research.

While hardly matters of life and death, concerns of a similar nature still apply for interventions seeking to shift energy consumers' behavior. Imagine one was testing the impact of normative messages indicating household electricity usage relative to similarly situated neighbors. Suppose one learns that the provision of this information, while reducing electricity consumption among those discovering their usage is above average, actually increases subsequent consumption (and bills) among those who find themselves below the norm (i.e., an unintended and undesirable consequence known as the 'boomerang effect'[22]). If this effect quickly becomes apparent, is one justified in continuing – for an extended period of monitoring – a trial that might serve to increase usage and expense among households with below-average electricity consumption (current 'under-consumers'), in pursuit of greater certainty about the intervention's impact and duration of effects; that is, in the interests of maximizing the scientific value and practical utility of the trial? Ethical challenges may arise over the course of any behaviour change study, so investigators must remain cognizant of relevant issues and ensure that ethical requirements are always upheld, even (potentially) at the expense of the scientific value that can be extracted from the study.

Practical and/or ethical constraints may also limit adoption of other recommendations we have outlined here. For example, sometimes it is simply not feasible to undertake random sampling and allocation in natural settings. In some circumstances, our unwavering pursuit of this ideal might see us deviate so far from realism we are left with a highly artificial testing environment that limits the external validity of our findings. Further, some forms of behavior do not lend themselves so easily to direct observation and objective measurement by third-parties, e.g., specific minute-by-minute energy-specific practices performed in the privacy of one's home are difficult to gauge from traditional metering. And while unobtrusive monitoring of a 'control group' that remains wholly unaware they are under observation greatly enhances internal validity – allowing strong causal claims to be made – this approach can be contentious from the point of view of 'informed consent', and ethically unsound in certain situations. This may be especially true to the extent one can argue that the behavior under observation is highly sensitive or private. Similarly, while we have recommended retaining and analyzing data on *all* those initially assigned to the treatment and control groups – irrespective of whether, in the end, they fully completed the program and furnished all records – we acknowledge the thorny issues around participant consent that inevitably arise with such 'intention to treat' analyses. Certainly it must be acknowledged and respected that participants who explicitly withdraw from a study (which is distinct from merely failing to follow through and fully comply with protocols) have the right also to withdraw all their data from analysis. As noted, in the face of such data loss, one is essentially left deriving then *not* the impact of the program per se, but rather the impact of fully participating in the program. Nevertheless, when scientific best practice fails to accord with ethical best practice, unquestionably it is the latter that must prevail. In sum, while RCTs are in principle the 'gold standard' for evaluating program impact, in practice we recognize there may be contexts and challenges that constrain one's willingness and/or capacity to adhere strictly to these principles and recommendations.

4. The real world of RCTs: energy-related behavioral interventions in practice

Despite these caveats, energy researchers, practitioners and policymakers need not be discouraged from capitalizing on the immense value of RCTs in real-world settings. There are many examples of randomized experiments being successfully implemented in practice. Certain types of behavior change strategies are ideal candidates for field-based RCTs, especially those that involve low- or no-cost variations to 'business-as-usual' processes, and where objective behavioral outcomes can be measured easily and inexpensively. For example, utilities already record the energy use of customers, so programs that target household energy consumption – via interventions delivered on business-as-usual quarterly bills – are always promising.[23] As Allcott and Mullainathan's [15,53] explain, when it comes to testing behavioral interventions in the field, randomization can be made practical via so-called 'encouragement designs' that concurrently evaluate a program's marketing (in terms of recruitment) and the program itself, e.g., that encourage a randomly-selected group to participate, but otherwise leave the program unaffected such that all customers (including controls) can still partake in the program. They also single out 'phased implementation' (e.g., randomized program phase-ins), which leverages the reality that organizations often roll-out new programs in a phased manner anyway, especially if resources are limited (see also Vine et al. [17] who discuss similar design practices, along with others, as being particularly appropriate for energy efficiency programs). Techniques such as these impose minimal extra burden on organizations, and in consequence are often less cumbersome and costly than one might imagine. Accordingly, focusing on those interventions that can be seamlessly integrated into business-as-usual processes at minimal cost is a good place to start—especially if a key priority is to identify cost-effective solutions that can be scaled to millions of consumers, which is often the case.

To illustrate, consider the various ways that governments and other organizations interact with consumers on a regular basis. In the context of energy conservation, governments may mandate disclosure of information to customers in the form of compulsory labeling schemes for products or services (e.g., appliance energy rating labels), and offer subsidies and advisory services to householders to promote energy efficiency home improvements (e.g., rebates for installing solar panels, insulation, etc.). And energy utilities send customers bill statements on a periodic basis that typically include tariff prices/rates and energy consumption information (akin to personal feedback). Utilities also may distribute marketing material to promote new energy-saving products/services, offer new tariff schemes, or solicit enrolment in demand-management programs. In all of these cases, the impact of information on consumer behavior can be heavily shaped by the way that information is presented. As such, governments and other organizations are ideally placed to experiment with different ways of designing, depicting and delivering the most impactful customer-focused

[22] For a field experiment examining the differential effects of social norms, see Schultz et al. [121] who found that descriptive normative information can have an unintended and undesirable effects on households consuming less energy than the neighborhood average.

[23] Moreover, as explained earlier, because RCTs require a sufficiently large number of 'units' to allow effective random assignment to treatment and control groups, individual- and household-level interventions are likely to be far easier to evaluate in practice than community-level programs.

communication. For instance, subtle variations in the framing of messages (e.g., using gain vs. loss-frames to promote energy efficiency; appealing to extrinsic vs. intrinsic motives), the layout of content (e.g., short and simple vs. complex and comprehensive), use of color, visuals and graphics (e.g., symbols, bar charts, infographics) and even the nature, frequency and mode of feedback (e.g., individual vs. comparative feedback; monthly vs. quarterly billing; postal vs. electronic mail) are all examples of things that could be experimentally tested in practice, using an organization's business-as-usual processes. It would be quite simple for an energy utility or appliance retailer to randomize across customers the content and layout of any communication to identify what format elicits the 'best' response. In the case of an energy study, the desired response could be anything from a reduction in overall or peak energy consumption, through to uptake of a new tariff offer, purchase of an energy efficient appliance, or enrolment in a demand management program or energy-saving initiative being promoted by the utility.

A program of experimental work undertaken by a US company called Opower – which constitutes one of the largest randomized field experiments in history – is an excellent example of how such strategies can be applied in practice [24,25,33,53]. In recent years, Opower has partnered with dozens of utilities across the country to test the impact of sending electricity customers 'home energy reports' on a regular basis (e.g., monthly, quarterly). The reports feature personalized energy use feedback, descriptive social norms (e.g., a graph that compares the household's energy use to that of similarly situated neighbors), as well as energy saving tips (for an overview of the program, see Ref. [24]). Using a robust experimental design, customers are randomly assigned either to receive a home energy report (experimental group) or not (control group), with the energy consumption of all households then monitored and compared over time.

Analyses of vast amounts of empirical data collected from hundreds of thousands of customers across the US suggest that these reports reduce electricity consumption in the average household by ~2% [24]. Allcott and Mullainathan [15] have estimated that a program like Opower costs a utility around 2.5¢ per kilowatt-hour saved and, if scaled nationwide across the US, could potentially reduce carbon dioxide emissions from electric power by 0.5%. More recently, Allcott and Rogers [33] found that even if reports are discontinued after two years, effects are relatively persistent (decaying at 10–20% per annum) and earlier estimates that assumed zero persistence had substantially understated the intervention's cost-effectiveness. Research of this nature provides robust evidence for the potential benefits of using non-pecuniary strategies to shift consumer behavior, and illustrates how relatively low-cost, high-impact solutions can be designed, tested and then cost-effectively scaled to millions of customers.[24] In recent years, a number of randomized field experiments have also been conducted to examine consumer response to time-varying electricity prices ('dynamic' or 'cost-reflective' pricing), often in conjunction with the deployment of advanced metering infrastructure and related technology (for a few examples, see Refs. [122–124]).

There are many other examples of relatively simple, easy and low-cost ways to embed unobtrusive experiments in real-world situations, and within business-as-usual processes, in order to rigorously evaluate the true impact of behavioral interventions at a population level. Policymakers, in particular, should keep this front of mind when designing new community-wide initiatives and policies to promote residential energy efficiency. For instance, if a government intends to use financial incentives to encourage citizens to purchase new energy efficient technology (e.g., solar and battery storage systems) or to participate in energy-saving initiatives (e.g., direct load control programs), it may be prudent to conduct a pilot trial to test whether there are differential participation rates when incentives of identical monetary value are framed in different ways, e.g., $200 presented either as a rebate, cash-back bonus, tax credit, or concession could potentially make a difference to consumer uptake/participation. Using a randomized experiment, policymakers could easily test whether an incentive framed as a $200 'rebate' is more appealing to consumers than a $200 'cash-back bonus' or $200 'discount'. One may even want to test how varying the monetary value of an incentive impacts consumer response, to identify the point of diminishing marginal returns, i.e., the exact threshold at which higher monetary values no longer yield proportionally higher returns. It may well be that consumers are generally motivated by any incentive, big or small, such that even low-value rewards produce the desired response, and anything beyond that is unnecessary expense. Again, it would be relatively straightforward to test these kinds of questions using a field experiment, gaining robust evidence regarding the relative cost-effectiveness, scalability and return on investment of alternative offers.[25] Future field experiments should continue to explore simple yet impactful means of promoting household energy efficiency and conservation, demand management and grid optimization. We should identify and test in real-world experimental trials a diverse array of behavioral strategies that, if proven impactful and cost-effective, can be easily scaled and successfully implemented en masse across the broader population of residential energy consumers.

5. Conclusions

Across diverse areas of public policy, behavior change interventions are now commonly deployed in an effort to shift people's behavior in desired directions—for example, toward healthier lifestyle choices, wiser financial decisions, and more environmentally-friendly practices. This extends to the specific domain of residential energy use, where a multitude of behavioral interventions and programs have been designed to shift the behavior of consumers and households in some desired way, e.g., toward greater energy efficiency, lower total and peak electricity usage, optimal responsiveness to dynamic tariffs, greater uptake of renewables and low-emission technology. Yet in many cases, the efficacy (thus cost-effectiveness) of such programs remains unknown, and indeed, unknowable. This may be due to fundamental limitations in program design, methodology and/or analysis, as we have explained here. Most notably, it is often the result of failing to build into the design of a behavior change program, from the very outset, the capacity to properly *evaluate* its success in a scientifically rigorous manner.

Against this backdrop, our paper has detailed some key principles for designing, conducting and evaluating the impact of such programs in the residential energy domain, where a multitude of behavioral strategies have been developed over the years in an effort to improve household energy efficiency and conservation, as well as participation in demand management, uptake of load control technologies, and grid optimization. We have described how designing an intervention in the form of an RCT greatly enhances our capacity to obtain accurate and readily generalizable estimates

[24] In 2014, Allcott and Rogers [33] reported that 6.2 million households served by 85 utilities across the US were receiving home energy reports.

[25] Note that steps would need to be taken to address any concerns over inequity or unfairness; and customers who received lower-value rebates may need to be compensated after the trial. By keeping the trial relatively small-scale, any household that was financially disadvantaged relative to others (e.g., those who accepted a lower-value rebate, or received no rebate at all) are reimbursed accordingly at the conclusion of the trial.

of a program's true impact on the outcome of interest, i.e., whether adoption of a particular strategy or exposure to a particular intervention actually *causes* a significant, measurable reduction in (say) household energy use. We have provided here a set of best practice guidelines for enhancing both the internal and external validity of claims about program impact, including specific recommendations for hypothesis formulation; program design; participant sampling, recruitment and allocation; program implementation; data collection; statistical analysis; and replication of results. By giving energy policymakers greater control over, and confidence in the scientific validity, reliability and generalizability of findings from behavior change programs, we hope that these guidelines serve to improve the cost-effectiveness and mass scalability of future energy behavior interventions across time, contexts and populations.

References

[1] P.W. Schultz, Strategies for promoting proenvironmental behavior: lots of tools but few instructions, Eur. Psychol. 19 (2014) 107–117.
[2] L. Steg, C. Vlek, Encouraging pro-environmental behaviour: an integrative review and research agenda, J. Environ. Psychol. 29 (2009) 309–317.
[3] L. Steg, Promoting household energy conservation, Energy Policy 36 (2008) 4449–4453.
[4] P. Stern, Information, incentives, and proenvironmental consumer behavior, J. Consumer Policy 22 (1999) 461–478.
[5] W. Abrahamse, L. Steg, C. Vlek, T. Rothengatter, A review of intervention studies aimed at household energy conservation, J. Environ. Psychol. 25 (2005) 273–291.
[6] M. Delmas, M. Fischlein, O.I. Asensio, Information strategies and energy conservation behavior: a meta-analysis of experimental studies from 1975 to 2012, Energy Policy 61 (2013) 729–739.
[7] W.O. Dwyer, F.C. Leeming, M.K. Cobern, B.E. Porter, J.M. Jackson, Critical review of behavioural interventions to preserve the environment: research since 1980, Environ. Behav. 25 (1993) 275–321.
[8] R. Osbaldiston, J.P. Schott, Environmental sustainability and behavioral science: a meta-analysis of proenvironmental behavior experiments, Environ. Behav. 44 (2012) 257–299.
[9] W. Abrahamse, L. Steg, Social influence approaches to encourage resource conservation: a meta-analysis, Global Environ. Change 23 (2013) 1773–1785.
[10] A.L. Davis, T. Krishnamurti, B. Fischhoff, W.B. de Bruin, Setting a standard for electricity pilot studies, Energy Policy 62 (2013) 401–409.
[11] K. Schulz, D. Altman, D. Moher, CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials, Ann. Intern. Med. 152 (2010) 726–732.
[12] P. Solomon, M.M. Cavanaugh, J. Draine, Randomized Controlled Trials: Design and Implementation for Community-Based Psychosocial Interventions, Oxford University Press, New York, US, 2009.
[13] J.M. Kendall, Designing a research project: randomised controlled trials and their principles, Emerg. Med. J. 20 (2003) 164–168.
[14] B. Sibbald, M. Roland, Understanding controlled trials: why are randomised controlled trials important? Br. Med. J. 316 (1998) 201.
[15] H. Allcott, S. Mullainathan, Behavior and energy policy, Science 327 (2010) 1204–1205.
[16] A. Todd, E. Stuart, C. Goldman, L. Berkley, S. Schiller, Evaluation, Measurement, and Verification (EM&V) for Behavior-Based Energy Efficiency Programs: Issues and Recommendations, ACEEE Summer Study on Energy Efficiency in Buildings, 2012.
[17] E. Vine, M. Sullivan, L. Lutzenhiser, C. Blumstein, B. Miller, Experimentation and the evaluation of energy efficiency programs, Energy Effic. 7 (4) (2014) 627–640, http://dx.doi.org/10.1007/s12053-013-9244-4.
[18] G. Gardner, P. Stern, Environmental Problems and Human Behavior, Pearson, Boston, 2002.
[19] P. Stern, G. Gardner, Psychological research and energy policy, Am. Psychol. 36 (1981) 329–342.
[20] I.M. Azevedo, Consumer end-use energy efficiency and rebound effects, Annu. Rev. Environ. Resour. 39 (2014) 393–418.
[21] A. Druckman, M. Chitnis, S. Sorrell, T. Jackson, Missing carbon reductions: exploring rebound and backfire effects in UK households, Energy Policy 39 (2011) 3572–3581.
[22] P.H. Berkhout, J.C. Muskens, J.W. Velthuijsen, Defining the rebound effect, Energy Policy 28 (2000) 425–432.
[23] L.A. Greening, D.L. Greene, C. Difiglio, Energy efficiency and consumption – the rebound effect – a survey, Energy Policy 28 (2000) 389–401.
[24] H. Allcott, Social norms and energy conservation, J. Public Econ. 95 (2011) 1082–1095.
[25] I. Ayres, S. Raseman, A. Shih, Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage, J. Law Econ. Org. 29 (2012) 992–1022.
[26] R.M.J. Benders, R. Kok, H.C. Moll, G. Wiersma, K.J. Noorman, New approaches for household energy conservation-in search of personal household energy budgets and energy reduction options, Energy Policy 34 (2006) 3612–3622.
[27] T.M. Harries, R. Rettie, M. Studley, K. Burchell, S. Chambers, Is social norms marketing effective: a case study in domestic electricity consumption, Eur. J. Market. 47 (2013) 1458–1475.
[28] J.M. Nolan, P.W. Schultz, R.B. Cialdini, N.J. Goldstein, V. Griskevicius, Normative social influence is underdetected, Pers. Soc. Psychol. Bull. 34 (2008) 913–923.
[29] B.K. Sovacool, What are we doing here? Analyzing fifteen years of energy scholarship and proposing a social science research agenda, Energy Res. Soc. Sci. 1 (2014) 1–29.
[30] B.K. Sovacool, Diversity: energy studies need social science, Nature 511 (2014) 529–530.
[31] A.R. Jadad, M.W. Enkin, Randomized Controlled Trials: Questions, Answers and Musings, Wiley-Blackwell, Chicester, 2007.
[32] W.F. Rosenberger, J.M. Lachin, Randomization in Clinical Trials: Theory and Practice, John Wiley & Sons, New York, 2002.
[33] H. Allcott, T. Rogers, The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation, Am. Econ. Rev. 104 (2014) 3003–3037.
[34] K.F. Punch, Introduction to Social Research: Quantitative and Qualitative Approaches, 3rd ed., Sage Publications Ltd., London, 2014.
[35] W.W. Charters Jr., Understanding Variables & Hypotheses in Scientific Research, ERIC Clearinghouse on Educational Management, University of Oregon, Eugene, OR, 1992.
[36] B.R. Flay, A. Biglan, R.F. Boruch, F.G. Castro, D. Gottfredson, S. Kellam, et al., Standards of evidence: criteria for efficacy, effectiveness and dissemination, Prev. Sci. 6 (2005) 151–175.
[37] M.L. Meldrum, A brief history of the randomized controlled trial: from oranges and lemons to the gold standard, Hematol. Oncol. Clin. North Am. 14 (2000) 745–760.
[38] K.W. Davidson, M. Goldstein, R.M. Kaplan, P.G. Kaufmann, G.L. Knatterud, C.T. Orleans, et al., Evidence-based behavioral medicine: what is it and how do we achieve it, Ann. Behav. Med. 26 (2003) 161–171.
[39] P. Armitage, The role of randomization in clinical trials, Stat. Med. 1 (1982) 345–352.
[40] D.B. Rubin, Causal inference using potential outcomes, J. Am. Stat. Assoc. 100 (2005) 322–331.
[41] D.B. Rubin, For objective causal inference, design trumps analysis, Ann. Appl. Stat. 2 (2008) 808–840.
[42] J.N. Druckman, D.P. Green, J.H. Kuklinski, A. Lupia, Cambridge Handbook of Experimental Political Science, Cambridge University Press, Cambridge, UK, 2011.
[43] M.B. Brewer, Research design and issues of validity, in: H.T. Reis, C.M. Judd (Eds.), Handbook of Research Methods in Social and Personality Psychology, Cambridge University Press, Cambridge, 2000, pp. 3–16 (Chapter 1).
[44] C. Winship, R.D. Mare, Models for sample selection bias, Annu. Rev. Sociol. 32 (1992) 7–50.
[45] M.A. Hernán, S. Hernández-Díaz, J.M. Robins, A structural approach to selection bias, Epidemiology 15 (2004) 615–625.
[46] B. Geddes, How the cases you choose affect the answers you get: selection bias in comparative politics, Polit. Anal. 2 (1990) 131–150.
[47] R.E. Larzelere, B.R. Kuhn, B. Johnson, The intervention selection bias: an underrecognized confound in intervention research, Psychol. Bull. 130 (2004) 289.
[48] D.G. Altman, J.M. Bland, Treatment allocation in controlled trials: why randomise? Br. Med. J. 318 (1999) 1209.
[49] D.G. Altman, Randomisation, Br. Med. J. 302 (1991) 1481–1482.
[50] R.S. Hartman, Self-selection bias in the evolution of voluntary energy conservation programs, Rev. Econ. Stat. 44 (1988) 8–58.
[51] P.W. Holland, Statistics and causal inference, J. Am. Stat. Assoc. 81 (1986) 945–960.
[52] D.A. Kenny, Correlation and Causality, Wiley, New York, 1979.
[53] H. Allcott, S. Mullainathan, Behavioral Science and Energy Policy, Massachusetts Institute of Technology, Cambridge, MA, USA, 2010, Working Paper (February).
[54] K. Schulz, D. Grimes, Sample size slippages in randomised trials: exclusions and the lost and wayward, Lancet 359 (2002) 781–785.
[55] K. Schulz, D. Grimes, Sample size calculations in randomised trials: mandatory and mystical, Lancet 365 (2005) 1348–1353.
[56] P. Dattalo, Determining Sample Size: Balancing Power, Precision, and Practicality, Oxford University Press, Oxford, 2008.
[57] T.P. Ryan, Sample Size Determination and Power, Wiley, 2013.
[58] R.V. Lenth, Some practical guidelines for effective sample size determination, Am. Stat. 55 (2001) 187–193.
[59] S.-C. Chow, H. Wang, J. Shao, Sample Size Calculations in Clinical Research, 2nd ed., Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL, 2007.
[60] D.A. Dzewaltowski, P.A. Estabrooks, L.M. Klesges, S. Bull, R.E. Glasgow, Behavior change intervention research in community settings: how generalizable are the results, Health Promot. Int. 19 (2004) 235–245.
[61] E. Aronson, P.C. Ellsworth, J.M. Carlsmith, M.H. Gonzales, Methods of Research in Social Psychology, McGraw-Hill Publishing Company, New York, 1990.
[62] P.M. Rothwell, External validity of randomised controlled trials: to whom do the results of this trial apply? Lancet 365 (2005) 82–93.

[63] A. Steckler, K.R. McLeroy, The importance of external validity, Am. J. Public Health 98 (2008) 9–10.
[64] R. Glasgow, L. Green, L. Klesges, D. Abrams, E. Fisher, M. Goldstein, et al., External validity: we need to do more, Ann. Behav. Med. 31 (2006) 105–108.
[65] S. Lohr, Sampling: Design and Analysis, 2nd ed., Books/Cole, Cengage Learning, Boston, MA, 2010.
[66] S. Treweek, P. Lockhart, M. Pitkethly, J.A. Cook, M. Kjeldstrøm, M. Johansen, et al., Methods to improve recruitment to randomised controlled trials: cochrane systematic review and meta-analysis, BMJ Open 3 (2013).
[67] J.M. Watson, D.J. Torgerson, Increasing recruitment to randomised trials: a review of randomised controlled trials, BMC Med. Res. Methodol. 6 (2006) 1–9.
[68] S. Treweek, E. Mitchell, M. Pitkethly, J. Cook, M. Kjeldstrøm, M. Johansen, T.K. Taskila, F. Sullivan, S. Wilson, C. Jackson, R. Jones, P. Lockhart, Strategies to improve recruitment to randomised controlled trials, Cochrane Database Syst. Rev. (2010), http://dx.doi.org/10.1002/14651858.MR000013.pub5, Issue 4. Art. No.: MR000013.
[69] M. Nelson, D. Cordray, C. Hulleman, C. Darrow, E. Sommer, A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions, J. Behav. Health Serv. Res. 39 (2012) 374–396.
[70] A.J. Bellg, B. Borrelli, B. Resnick, J. Hecht, D.S. Minicucci, M. Ory, et al., Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium, Health Psychol. 23 (2004) 443.
[71] F.J. Moncher, R.J. Prinz, Treatment fidelity in outcome studies, Clin. Psychol. Rev. 11 (1991) 247–266.
[72] B. Borrelli, D. Sepinwall, D. Ernst, A.J. Bellg, S. Czajkowski, R. Breger, et al., A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research, J. Consult. Clin. Psychol. 73 (2005) 852.
[73] D.T. Campbell, J.C. Stanley, Experimental and Quasi-experimental Designs for Research, Rand McNally, Chicago, 1963.
[74] J.J. McGonigle, J. Rojahn, J. Dixon, P.S. Strain, Multiple treatment interference in the alternating treatments design as a function of the intercomponent interval length, J. Appl. Behav. Anal. 20 (1987) 171–178.
[75] E.S. Shapiro, A.E. Kazdin, J.J. McGonigle, Multiple-treatment interference in the simultaneous-or alternating-treatments design, Behav. Assess. 4 (1982) 105–115.
[76] B. Karlin, J.F. Zinger, R. Ford, The effects of feedback on energy conservation: a meta-analysis, Psychol. Bull. 141 (2015) 1205–1227.
[77] S. Donaldson, E. Grant-Vallone, Understanding self-report bias in organizational behavior research, J. Bus. Psychol. 17 (2002) 245–260.
[78] R.H. Moorman, P.M. Podsakoff, A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research, J. Occup. Org. Psychol. 65 (1992) 131–149.
[79] M.F. King, G.C. Bruner, Social desirability bias: a neglected aspect of validity testing, Psychol. Market. 17 (2000) 79–103.
[80] D.L. Paulhus, Socially desirable responding: the evolution of a construct, in: H.I. Braun, D.N. Jackson, D.E. Wiley (Eds.), The Role of Constructs in Psychological and Educational Measurement, Lawrence Erlbaum Associates, Inc., New Jersey, 2002, pp. 51–76.
[81] T. Holtgraves, Social desirability and self-reports: testing models of socially desirable responding, Pers. Soc. Psychol. Bull. 30 (2004) 161–172.
[82] N. Schwarz, Self-reports: how the questions shape the answers, Am. Psychol. 54 (1999) 93.
[83] L. Bickman, D.J. Rog, The SAGE Handbook of Applied Social Research Methods, 2nd ed., Sage Publications, Thousand Oaks, CA, 2008.
[84] J.M. Lachin, The role of measurement reliability in clinical trials, Clin. Trials 1 (2004) 553–566.
[85] R. Adcock, Measurement validity: a shared standard for qualitative and quantitative research, Am. Polit. Sci. Rev. 95 (2001) 529–546.
[86] L. Wilkinson, Statistical methods in psychology journals: guidelines and explanations, Am. Psychol. 54 (1999) 594.
[87] E.J. Webb, D.T. Campbell, R.D. Schwartz, L. Sechrest, Unobtrusive Measures: Nonreactive Research in the Social Sciences, Rand McNally, Chicago, 1966.
[88] A.E. Kazdin, Unobtrusive measures in behavioral assessment, J. Appl. Behav. Anal. 12 (1979) 713–724.
[89] M.T. Orne, On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications, Am. Psychol. 17 (1962) 776.
[90] A.L. Nichols, J.K. Maner, The good-subject effect: investigating participant demand characteristics, J. Gen. Psychol. 135 (2008) 151–165.
[91] J. McCambridge, M. De Bruin, J. Witton, The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review, PLoS One 7 (2012) e39116.
[92] R. Rosenthal, R.L. Rosnow, Artifacts in Behavioral Research, Oxford University Press, Oxford, 2009.
[93] M.R. Leary, R.M. Kowalski, Impression management: a literature review and two-component model, Psychol. Bull. 107 (1990) 34.
[94] R.H. Franke, J.D. Kaul, The Hawthorne experiments: first statistical interpretation, Am. Sociol. Rev. 43 (1978) 623–643.
[95] J.G. Adair, The Hawthorne effect: a reconsideration of the methodological artifact, J. Appl. Psychol. 69 (1984) 334.
[96] P. Courtenay-Hall, L. Rogers, Gaps in mind: problems in environmental knowledge-behaviour modelling research, Environ. Educ. Res. 8 (2002) 283–297.
[97] T. Kennedy, G. Regehr, J. Rosenfield, S.W. Roberts, L. Lingard, Exploring the gap between knowledge and behavior: a qualitative study of clinician action following an educational intervention, Acad. Med. 79 (2004) 386–393.
[98] A. Kollmuss, J. Agyeman, Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior, Environ. Educ. Res. 8 (2002) 239–260.
[99] F.X. Sligo, A.M. Jameson, The knowledge-behavior gap in use of health information, J. Am. Soc. Inf. Sci. 51 (2000) 858–869.
[100] J. Blake, Overcoming the 'value-action gap' in environmental policy: tensions between national policy and local experience, Int. J. Justice Sustain. 4 (1999) 257–278.
[101] E. Boulstridge, M. Carrigan, Do consumers really care about corporate responsibility? Highlighting the attitude-behaviour gap, J. Commun. Manage. 4 (2000) 355–368.
[102] R. Flynn, P. Bellaby, M. Ricci, The 'value-action gap' in public attitudes towards sustainable energy: the case of hydrogen energy, Sociol. Rev. 57 (2010) 159–180.
[103] E. Huddart-Kennedy, T.M. Beckley, B.L. McFarlane, S. Nadeau, Why we don't walk the talk: understanding the environmental values/behaviour gap in Canada, Hum. Ecol. Rev. 16 (2009) 151–160.
[104] P. Sheeran, C. Abraham, Mediator of moderators: temporal stability of intention and the intention-behavior relation, Pers. Soc. Psychol. Bull. 29 (2003) 205–215.
[105] P. Sheeran, Intention-behavior relations: a conceptual and empirical review, Eur. Rev. Soc. Psychol. 12 (2002) 1–36.
[106] B.M. Fennis, M.A. Adriaanse, W. Stroebe, B. Pol, Bridging the intention–behavior gap: inducing implementation intentions through persuasive appeals, J. Consum. Psychol. 21 (2011) 302–311.
[107] T.L. Webb, P. Sheeran, Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence, Psychol. Bull. 132 (2006) 249–268.
[108] I.B. Weiner, J.A. Schinka, W.F. Velicer, Handbook of Psychology, Research Methods in Psychology, 2nd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.
[109] T. Hinkin, A brief tutorial on the development of measures for use in survey questionnaires, Org. Res. Methods 1 (1998) 104–121.
[110] T. Hinkin, A review of scale development practices in the study of organizations, J. Manage. 21 (1995) 967–988.
[111] D.P. Schwab, Research Methods for Organizational Studies, 2nd ed., Psychology Press, New York, 2005.
[112] H. Allcott, T. Rogers, How long do treatment effects last? Persistence and durability of a descriptive norms intervention's effect on energy conservation, HKS Faculty Research Working Paper (2012) RWP12-045, 1–35.
[113] K.S. Fielding, A. Spinks, S. Russell, R. McCrea, R. Stewart, J. Gardner, An experimental test of voluntary strategies to promote urban water demand management, J. Environ. Manage. 114 (2013) 343–351.
[114] K. Kelley, K.J. Preacher, On effect size, Psychol. Methods 17 (2012) 137–152.
[115] T. Baguley, Standardized or simple effect size: what should be reported, Br. J. Psychol. 100 (2009) 603–617.
[116] S. Michie, C. Abraham, Interventions to change health behaviours: evidence-based or evidence-inspired? Psychol. Health 19 (2004) 29–49.
[117] S.W. Raudenbush, Comparing personal trajectories and drawing causal inferences from longitudinal data, Annu. Rev. Psychol. 52 (2001) 501–525.
[118] S. Hollis, F. Campbell, What is meant by intention to treat analysis: survey of published randomised controlled trials, BMJ 319 (1999) 670–674.
[119] J.P. Higgins, S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0, The Cochrane Collaboration. Available from: www.handbook.cochrane.org, 2011 (updated March 2011).
[120] S. Edwards, R.J. Lilford, D. Braunholtz, J.M. Jackson, J. Hewison, J. Thornton, Ethical issues in the design and conduct of randomised controlled trials, Health Technol. Assess. 2 (1998) 1–146.
[121] P.W. Schultz, J.M. Nolan, R.B. Cialdini, N.J. Goldstein, V. Griskevicius, The constructive, destructive, and reconstructive power of social norms, Psychol. Sci. 18 (2007) 429–434.
[122] A. Faruqui, S. Sergici, Dynamic pricing of electricity in the mid-atlantic region: econometric results from the Baltimore gas and electric company experiment, J. Regul. Econ. 40 (2011) 82–109.
[123] T. Ida, K. Ito, M. Tanaka, Using dynamic electricity pricing to address energy crises: evidence from randomized field experiments, in: 36th Annual NBER Summer Institute, Cambridge, MA, USA, 2013, Available online at: http://www.econ.kyoto-u.ac.jp/~ida/4Hoka/smagri/20133023Ito_Ida_Tanaka_Dynamic_Pricing.pdf.
[124] A. Todd, P. Cappers, C. Goldman, Residential Customer Enrollment in Time-based Rate and Enabling Technology Programs, Lawrence Berkeley National Laboratory, 2013.