Energy Research & Social Science 59 (2020) 101260

Contents lists available at ScienceDirect

Energy Research & Social Science

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

Original research article

Ensuring statistics have power: Guidance for designing, reporting and acting on electricity demand reduction and behaviour change programs

Ben Anderson[a,b,*], Tom Rushby[a], Abubakr Bahaj[a], Patrick James[a]

[a] Energy and Climate Change Division/Sustainable Energy Research Group, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom

[b] Centre for Sustainability, University of Otago, PO Box 56, Dunedin 9054, New Zealand

ARTICLE INFO

Keywords:
Statistical power
Statistical significance
Energy studies
Study design
Sample size

ABSTRACT

In this paper we address ongoing confusion over the meaning of *statistical significance* and *statistical power* in energy efficiency and energy demand reduction intervention studies. We discuss the role of these concepts in designing studies, in deciding what can be inferred from the results and thus what course of subsequent action to take. We do this using a worked example of a study of Heat Pump demand response in New Zealand to show how to appropriately size experimental and observational studies, the consequences this has for subsequent data analysis and the decisions that can then be taken. The paper then provides two sets of recommendations. The first focuses on the uncontroversial but seemingly ignorable issue of statistical power analysis and sample design, something regularly omitted in the energy studies literature. The second focuses on how to report energy demand reduction study or trial results, make inferences and take commercial or policy-oriented decisions in a contextually appropriate way. The paper therefore offers guidance to researchers tasked with designing and assessing such studies; project managers who need to understand what can count as evidence, for what purpose and in what context and decision makers who need to make defensible commercial or policy decisions based on that evidence. The paper therefore helps all of these stakeholders to distinguish the search for statistical significance from the requirement for actionable evidence and so avoid throwing the substantive baby out with the *p*-value bathwater.

1. Introduction

The ongoing debate over the relative merits of *p*-value based significance testing and inference across a range of scientific fields [2,15,31] resonates with our experiences of designing and running empirical energy efficiency and energy demand studies. Whether experimental or observational, we have noted ongoing confusion amongst researchers involved in these studies over the meaning, value and use of two key statistical concepts: *statistical significance* and *statistical power*. This is compounded by confusion over how these concepts should be used both in *designing* studies and also in deciding what can be *inferred* from the results and thus what course of subsequent action to take.

Echoing recent commentary on the quality of energy related social science papers [25], we have found this to be the case in academic research where the objective is to uncover 'the most likely explanation' under established academic conventions. However, and what prompted this paper, we have also found it to be the case in applied research where the objective is to 'make a robust and defensible decision' for commercial strategy or public policy purposes based on the balance of evidence and probability.

We have observed three consequences of this confusion: Firstly, as has been reported in other fields [20,27], a large number of energy studies have been implemented with no real idea of whether they will be able to robustly test their hypotheses under normal academic conventions [13,25]. Secondly, these and other studies which *have* been more robustly designed risk being dismissed and/or themselves dismissing potentially useful results as not being statistically significant due to a very narrow application of p-value based statistical significance testing. Finally, a lack of consistency of reporting, particularly in applied research, makes comparing across studies and thus developing a synthesised and summative evidence base for strategic or public policy decision making extremely difficult.

In this brief paper we respond to these confusions using a worked example: the design of a hypothetical household electricity demand

* Corresponding author at: Energy and Climate Change Division/Sustainable Energy Research Group, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom.
E-mail address: b.anderson@soton.ac.uk (B. Anderson).

https://doi.org/10.1016/j.erss.2019.101260
Received 17 April 2019; Received in revised form 26 June 2019; Accepted 15 August 2019

study in New Zealand which seeks to understand the patterns of Heat Pump use in different kinds of households in the evening winter peak demand period. We use this example to explain and demonstrate the role of statistical power and statistical significance in sample design, testing for differences (or intervention effects) and subsequent decision making whether in the commercial or public sector.

The paper is intended to offer guidance to energy sector professionals tasked with designing and assessing such trials; project managers who may not need to know the detail of our arguments but need to understand what can count as evidence, for what purpose and in what context, and decision makers who need to make defensible commercial or policy intervention decisions based on that evidence. The paper can therefore help all of these stakeholders to distinguish the search for statistical significance from the need for actionable evidence.

2. Error, power, significance and decision making

Two types of error are of concern in both academic and applied energy research – Type I (false positives) and Type II (false negatives) [8,9].

Type I: a false positive - an effect is inferred when in fact there is none. The significance level (p value) of the statistical test to be used to assess the efficacy of an intervention represents both the extent to which the observed data matches the null model to be tested [31], and also the risk of a Type I error. In most trials the null model will be a measure of 'no difference' between control and intervention groups. By convention, the p value *threshold* for rejecting the null model (the risk of a Type I error) is generally set to 0.05 (5%) although this choice is entirely subjective and reflects human perceptions of what constitutes an unlikely event. In this instance, 5% (or 1 in 20) is considered to represent an unlikely event and generally derives from the medical sciences where the impact of a false positive may be both actively harmful and very costly. In commercial or policy terms an action taken on a result with a larger p value (e.g. setting the p value threshold to 10%) would increase the risk of making a Type I error and thus implementing a potentially costly intervention that is unlikely to have the effect desired. However, as we discuss in more detail below, this is *not necessarily* bad practice, where there is a low opportunity cost in terms of capital, equipment or customer contact and education, little risk of harm and some chance of benefit to the stakeholders involved;

Type II: a false negative - an effect is not inferred when in fact there is one. The pre-study risk of making a Type II error is known as statistical power [7,15,19] and is normally set to 0.8 (80%) by convention. Again, this generally derives from the medical sciences where a risk of a false negative has associated likelihood of direct harm through avoidable deaths and potential legal liabilities. From a commercial or policy perspective reducing power (e.g. to 0.7 or 70%) will therefore increase the risk of making a Type II error and so taking no action when in fact the intervention would probably have had the effect desired. As above, in the case of energy system interventions a balance of probabilities such as 60:40 may be a sufficient risk (rather than 80:20) where only a limited commercial benefit would flow from action, the costs of that action are high and the likelihood of disbenefit in the case of non-intervention is low.

Clearly in both cases one might expect to set the relevant threshold higher if specific vulnerable groups are to be targeted for whom the consequences of action (or inaction) may be more beneficial (or detrimental).

Statistical power calculations enable the investigator to estimate the sample size that would be needed to robustly detect an experimental effect with a given risk of a false positive (Type I error) or false negative (Type II error) result. This prevents a study from recruiting too few participants to be able to robustly detect the hypothesised intervention effect [10] or wasting resources by recruiting a larger sample than needed [19].

Previous work has suggested that sample sizes in most energy efficiency studies may be too low to provide adequate power and so statistically robust conclusions cannot be drawn at conventional thresholds [13]. A more recent review focusing on demand response studies reached a similar conclusion [26]. It is therefore hardly surprising that a number of studies report effect sizes which are not statistically significant at conventional thresholds [26], choose to use lower statistical significance thresholds [1,5,23,24] or lower both statistical power values *and* statistical significance thresholds [24,28,29].

However it would be wrong to conclude that this is *necessarily* bad practice. Both historical [8] and more recent discussions of the role of p values in inference [15,17,31] and recent commentary on the replicability crisis in both energy studies and elsewhere [4,16,18,21], remind us that decisions should never be based solely on statistical significance thresholds set purely by convention. Rather, inference and thus decision making should be based on an assessment of *all* of the following:

- difference or effect size - is it 2% or 22% (i.e. is the result *important* or *useful*, "What is the estimated *bang for buck*?");
- statistical confidence intervals - (i.e. is there *uncertainty* or *variation* in response, "How uncertain is the estimated bang?");
- statistical p values - (i.e. what is the risk of a Type I error / *false positive*, "What is the risk the bang observed isn't real?") and;
- statistical power - (i.e. what is the risk of a Type II error / *false negative*, "What is the risk there is a bang when we concluded there wasn't?")

Only then can a contextually appropriate decision be taken as to whether the effect is large enough, certain enough and has a low enough risk of being a false positive or false negative to warrant action. In doing so one can avoid simply dismissing results on the basis of a failure to meet conventional statistical levels of significance.

In the following sections we apply these principles to the design of a hypothetical New Zealand household electricity demand study focused on winter evening heat pump demand. We then apply them to the use of a simple statistical test of difference between household types to demonstrate and clarify these points.

3. Sample design: statistical power

To return to the discussion of statistical power, we need to estimate the size of the groups we will require. This is crucial to resource budgeting ("*How many households and thus $ do I need?*") and ensuring good study design practice ("*Will I be able to answer my research question?*") [13,25]. In both cases the answer is not absolute since it will depend on our tolerance of Type I and Type II error risks.

Calculation of the required sample group sizes requires the estimation of the probable size of the difference or intervention effect, agreement on the significance level of the statistical test to be used (p value threshold or Type I error risk) and agreement on the level of statistical power (Type II error risk). Given any three of these values the fourth can be calculated if we also have an estimate of the mean and standard deviation of a numeric outcome to be measured. However it is worth noting that Type I and Type II errors trade off against each other such that reducing one will automatically increase the other unless the sample size is increased.

In the case of demand response interventions, the effect size comprises a given % reduction in energy demand in a given time period. In the case of an analysis of difference, the effect size comprises the difference in energy demand between the groups under study. Estimates of the expected reduction (or difference) and thus the likely 'effect size', as well as the mean and variance can be derived from previous studies or data. In doing so, recent work recommends that analysis focuses on designing samples to robustly detect the smallest effect size that would make the intervention (cost) effective in its context [14]. Thus, if a business case demands a 10% effect, the samples should be designed with this in mind since any larger effect size detected will simply

improve the business case but any smaller value would mean the business case falls.

As we have noted, the choice of significance level (p value threshold) and statistical power are subjective and normative. Most academic researchers will struggle to justify deviating from the conventional power = 0.8 and $p = 0.01$ (or the slightly less conservative $p = 0.05$) values generally expected by academic journal editors and their reviewers. After all, they will need to publish their results and, notwithstanding current debates, these are the 'normal' quality-of-evidence indicators for the publication of academic journal articles.

However, as we have discussed there may be good reasons in applied research to take action on results of studies that use less conservative thresholds. These might include small-n studies that show large mean effect sizes for a low cost intervention but with sufficient uncertainty (large p values) to increase the risk of Type I error. In this case the low cost of the intervention balanced against the potentially large effect might make implementation worth the risk.

Nevertheless there is a strong argument for designing such studies using the more conservative conventional levels [13,25]. This enables them to satisfy 'academic norms' and *also* allows the application of a more relaxed approach to Type I or Type II error risks than is considered 'normal' in academic research if effect sizes or eventual sample sizes are smaller than expected.

To provide a worked example we have used winter 2015 weekday peak period (16:00–20:00) 'Heat Pump' electricity demand data extracted from the publicly available New Zealand Green Grid household electricity demand dataset [3] to conduct a set of statistical power analyses. The data comprises circuit level extracts from the 1 min level multi-circuit data for all 47 households.[1] The final dataset comprises data for 28 heat-pump owning households and the overall mean and standard deviation for power demand is shown in Table 1.[2] These values are the starting point for our statistical power analysis.[3]

The analysis was conducted using using knitr [34] in RStudio with R version 3.5.1 (2018-07-02) running on x86_64-apple-darwin15.6.0. Data manipulation, analysis and visualisation used data.table [11], dplyr [33] and ggplot2 [32]. Data summaries and t tests were conducted using base statistics [22] while power analysis was conducted using power.t.test [22] and pwr::pwr.2p.test [6].

Initially, we have used the values shown in Table 1 to calculate the intervention effect sizes (or magnitudes of difference) that could be robustly detected with power = 0.8, and the conventional and conservative $p = 0.01$.

The results are shown in Fig. 1 and we have included a reference line for a sample size of 1000 which would be able to detect an effect size of 9.08%. This means that if we expected there to be a difference in demand between household types of 9.29% and we wished to use a the conservative $p = 0.01$ threshold in analysis, we would need 1000 households in *each* of our household types. Similarly, if we were designing an intervention trial and we expected our intervention to achieve a 9% effect, if we wished to use the conventional p = 0.01 threshold in analysis we would need 1000 households in *each* of our trial groups.

We have also calculated the same effect size values for other p value thresholds (see Fig. 2 and Table 2). Were a study to be less risk averse in its inference and decision making but still wished to conform to academic norms then $p = 0.05$ may be acceptable. In this case, Fig. 2 and Table 2 show that only 600 households would be needed in each group.

On the other hand, we may decide that $p = 0.1$ is acceptable in which case only 425 households would be needed in each group.

Table 1
Summary electricity power demand statistics for heat pump owners during the winter weekday evening peak period (NZ GREEN Grid Household Electricity Demand data).

Season	N households	Mean W	SD W
Winter	27	410.65	269.15

However, in both cases the risk of a Type I error would increase from a 1 in 100 probability to a 1 in 20 and 1 in 10 probability respectively.

Further calculations could also be conducted with less conservative power values (e.g. 0.7). This will also decrease the number of households required but would increase the risk of a Type II error as noted above.

Not all studies are interested in testing the difference or change in a consumption variable. They may instead be interested in the prevalence of some attribute in a population and the extent to which this changes (or differs) between two groups. Calculating required sample sizes in this case follows a similar process but depends on the relative proportions of the attributes expected and the acceptable margin of error. A useful discussion of these calculations together with an applicable table of results can be found in Sovacool et al [25] (Table 5).

Suppose, for illustrative purposes we were interested in the adoption of heat pumps in two equal sized samples. Previous data might suggest that in one sample (say, home owners) the rate might be 40% and in rental properties it might be 25% [30]. Table 3 shows the sample size required to conclude a significant difference with power = 0.8 and at various p values.

We can repeat this for other proportions. For example, suppose both were much smaller and the magnitude of the difference was also smaller (e.g. 10% and 15%). Clearly, as Table 4 shows, we need much larger samples.

The above used an arcsine transform as recommended in [7] and implemented in [6] and could be repeated for any estimated proportions.

Alternatively, we may be interested in establishing the likely margins of error around a proportion given a certain sample size. To do this we can use Eq. (1) below.

$$me = \pm z^* \sqrt{\frac{p(1-p)}{n-1}} \quad (1)$$

Where me is the margin of error, p is the sample proportion, n is the sample size, and z is the appropriate z score value for the Type II error risk (p value) to be tolerated.

As an example, to re-use the results of Table 3, if we assume that the proportion we are interested in is 0.4 (40%) and our sample size is 151, then the above equation enables us to calculate the margin of error as +/- 0.078 (7.8%) with $p = 0.05$ ($z = 1.96$). Thus we could quote the Heat Pump uptake for owner-occupiers as 40% (+/- 7.8% [i.e. 32.2% – 47.8%] with $p = 0.05$.

However, this may be far too wide an error margin for our purposes and because we calculated this before we conducted our full study, we may instead have decided to recruit 500 per sample. In this case the margin of error is +/- 0.043 (4.3%) with $p = 0.05$ ($z = 1.96$) so we can now quote the Heat Pump uptake for owner-occupiers as 40% (+/- 4.3% [i.e. 35.7% – 44.3%] with $p = 0.05$. Increasing the sample size has increased the precision of our estimate and will also have increased the likelihood that we are comfortable concluding there are significant differences between home owners and renters (for example).

We now have enough information to recommend a sample size for our hypothetical trial or study. If we were to pursue a heat pump intervention trial then we would also require random allocation to a control and three intervention groups [13] as we want to test three different demand response interventions. On the other hand we may just be interested in the difference between different types of

[1] See https://github.com/CfSOtago/GREENGridData/

[2] -ve W values have been removed as they are considered to be rare instrument errors (see https://github.com/CfSOtago/GREENGridData/issues/6)

[3] Full code for the analysis contained in this paper is available from https://git.soton.ac.uk/ba1e12/weGotThePower

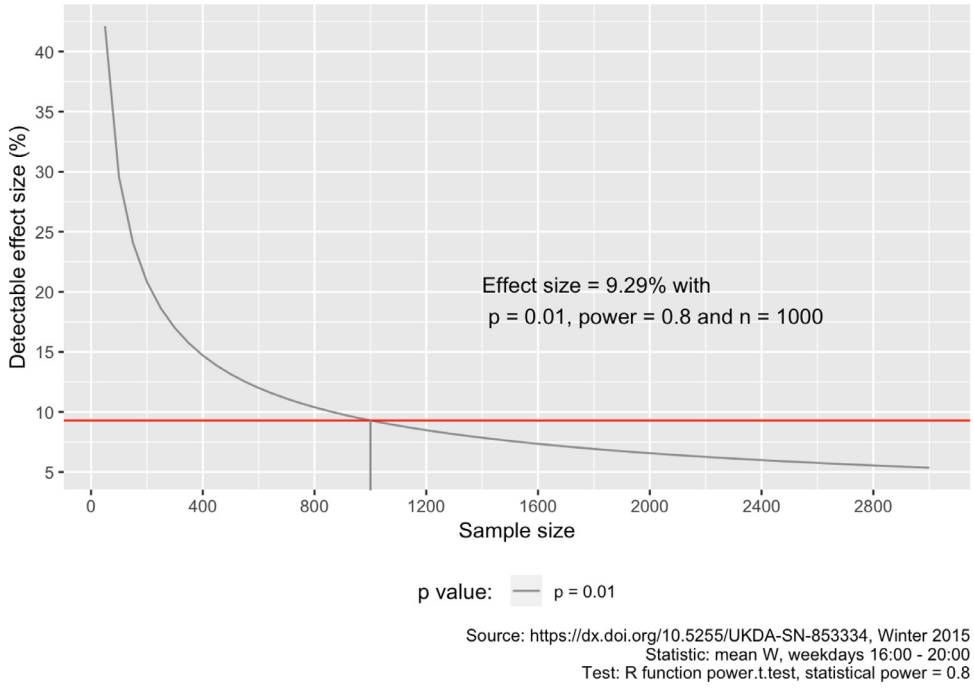

Fig. 1. Power analysis results (power = 0.8, p = 0.01).

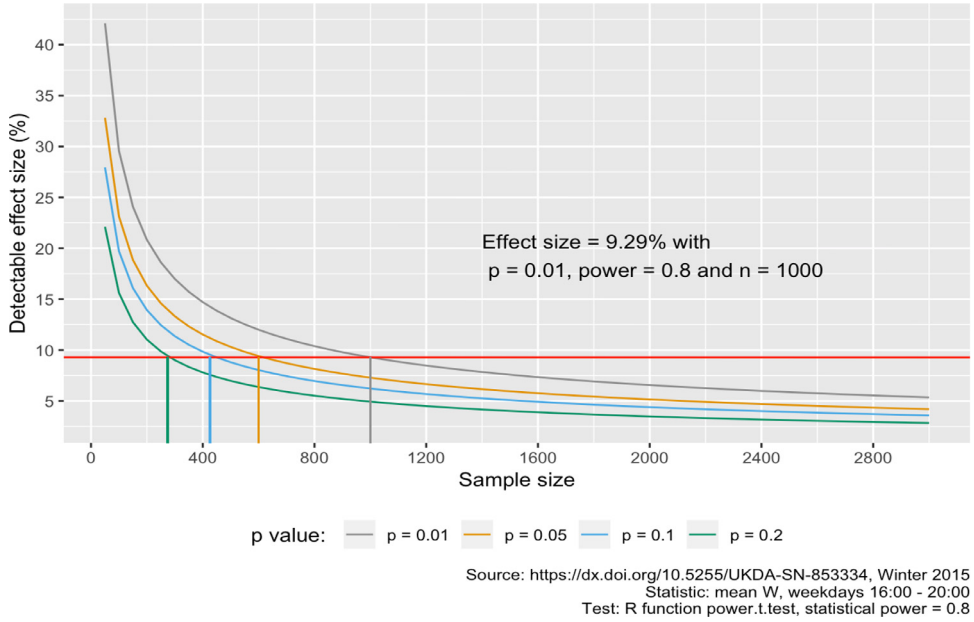

Fig. 2. Full power analysis results (power = 0.8, p = 0.01, 0.05, 0.1, 0.2).

household. It really does not matter which is the case – the same principles apply. In the next section we imagine that a study has been designed and implemented and that we now wish to conduct an analysis of the difference in Heat Pump demand between different household types.

4. Statistical significance: 'effect' sizes, confidence intervals and *p* values

4.1. Getting it 'wrong'

Let us imagine that we have ignored the power calculations reported in the previous section (e.g. Table 2) and decided, perhaps for cost reasons to recruit ~ 50 households in total. Now we wish to test for differences between households of different types.

Table 5 shows the mean and standard deviation for the winter heat pump power demand of four household types defined by the number of occupants using the winter GREEN Grid Heat Pump power demand data [3]. To do this we have created a sample of 52 households based on the 27 described in Table 1 using random re-sampling with replacement.

As we would expect, there are substantial differences between the groups and also notable differences in their frequency in the data reflecting the original sampling approach.[4]

[4] A focus on multi-person households during the snowball recruitment produced a biased sample that was not representative of New Zealand households making the generalisation of results to the New Zealand population

Table 2
Full power analysis results (power = 0.8, sample $N <= 1000$).

	Effect size			
Sample N	*p = 0.01*	*p = 0.05*	*p = 0.1*	*p = 0.2*
50	42.11	32.82	27.95	22.10
100	29.57	23.13	19.72	15.62
150	24.09	18.86	16.09	12.75
200	20.83	16.32	13.93	11.04
250	18.62	14.60	12.46	9.87
300	16.99	13.32	11.37	9.01
350	15.73	12.33	10.53	8.34
400	14.71	11.53	9.85	7.80
450	13.86	10.87	9.28	7.36
500	13.15	10.31	8.80	6.98
550	12.54	9.83	8.39	6.65
600	12.00	9.41	8.04	6.37
650	11.53	9.04	7.72	6.12
700	11.11	8.72	7.44	5.90
750	10.73	8.42	7.19	5.70
800	10.39	8.15	6.96	5.52
850	10.08	7.91	6.75	5.35
900	9.80	7.69	6.56	5.20
950	9.53	7.48	6.39	5.06
1000	9.29	7.29	6.22	4.93

Table 3
Samples required if proportion 1 = 40% and proportion 2 = 25%.

Significance level (*p* value)	Power	*n*
0.01	0.8	225
0.05	0.8	151
0.10	0.8	119
0.20	0.8	87

Table 4
Samples required if proportion 1 = 10% and proportion 2 = 15%.

Significance level (*p* value)	Power	*n*
0.01	0.8	1012
0.05	0.8	680
0.10	0.8	536
0.20	0.8	390

Table 5
Number of households and summary statistics per group.

N people	N households	mean W	SD W
1	3	219.94	174.63
2	6	273.15	60.40
3	20	387.96	229.45
4+	23	440.58	276.30

Next, we plot the differences using the mean and 95% confidence intervals (Fig. 3). As we would expect single person households appear to have much lower Heat Pump demand than the 3 and 4+ person households but the error bars indicate the uncertainty (variation) around the mean for each group. Based on this, we suspect that we may see low *p* values when we use statistical tests of the differences as the error bars overlap in most cases.

A t-test of the difference between the '1 person' and '3 person' groups produces the result shown in Table 6.

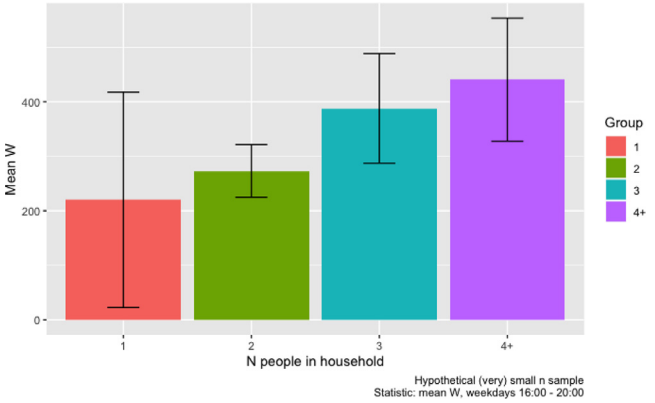

Fig. 3. Mean power demand by household type (Error bars = 95% confidence intervals for the mean).

Table 6
T-test results (1 person vs 3 person households).

1 person mean	3 persons mean	Mean difference	statistic	*P* value	95% CI (lower)	95% CI (upper)
219.94	387.96	–168.013	–1.485	0.230	–518.67	182.64

The results show that the mean difference in power demand between the groups was 168 W. This is a (very) large difference in the mean. The results of the t test show:

- difference (effect size) = 168 W or 43%;
- 95% confidence interval for the test = -518.67 to 182.64 representing *very substantial* uncertainty/variation;
- *p* value of 0.230 representing a rather *high* risk of a false positive result which fails all conventional *p* value thresholds as well as $p = 0.2$.

Thus, whilst there appears to be a large difference in evening peak winter heat pump power demand between these two household types, we cannot conclude that the difference is statistically significant at any conventional *p* value threshold.

Had this been a trial of three different interventions compared to a control, what would we have concluded? We have a large effect size, substantial uncertainty and an elevated risk of a false positive or Type I error (large *p* value) when compared to conventional *p* value levels. From a narrow and conventional '*p* value testing' perspective we would have concluded that there was no statistically significant difference between the Control and the Intervention Group. However this misses the crucial point that an organisation with a higher risk tolerance might conclude that the large effect size justifies implementing the intervention *even though* the risk of a false positive ($p = 0.230$) is much higher than the conventional thresholds.

But what about the difference between 1 person households and 4+ person households? In this case the t-test results are slightly different (Table 7).

Now:

- difference (effect size) = 220 W or 50%;
- 95% confidence interval for the test = −562 0.71 to 121.44 representing *substantial* uncertainty/variation;
- *p* value of 0.140 representing a relatively *low* risk of a false positive result but which fails the conventional $p < 0.05$ threshold and also the $p < 0.1$ threshold.

As before, despite the large W difference, we cannot conclude that the difference is statistically significant at any conventional *p* value

(*footnote continued*)
problematic. However, whilst crucial to making valid generalisations to populations of interest [25], a detailed discussion of sample bias is outside the scope of this paper.

Table 7
T test results (1 person vs 4+ person households).

1 person mean	4+ persons mean	Mean difference	statistic	*P* value	95% CI (lower)	95% CI (upper)
219.94	440.580	–220.640	–1.900	0.140	–562.713	121.437

threshold. Again, had this been an intervention trial then the subsequent action we would take would depend on our tolerance of Type I Error (false positive) risk. We have a substantial (and larger) effect size and we can be more certain about it as we have a lower risk of a Type I error (smaller *p* value). Implementing 'Intervention 2' would therefore seems a safer bet (or at least a more defensible risk).

It should be clear that in both cases our decision-making is somewhat hampered by the small sample size even though we have extremely large effect sizes. Perhaps we should have recruited rather more households.

4.2. Getting it 'right'

Suppose instead that we had designed and implemented our sample recruitment according to Table 2 so that we have a reasonable chance of detecting a difference of ~ 30% between the household types with power = 0.8 and at a significance level (*p*) of 0.05. This means we should have a sample of at least 100 in each of the four household types.

Table 8 shows the results for such a sample which has been derived by randomly re-sampling with replacement from the small sample of heat pump equipped households [3] reported in Table 1 with the aim of achieving at least ~100 in each cell. Due to the rarity of 1 person households this resampling has the effect of inflating the frequency of larger households. As we would expect the means, standard deviations and magnitude of the differences between the groups are similar although there are random fluctuations.

However, in comparison to Fig. 3 we can now see that the 95% confidence intervals for the group means are much narrower (Fig. 4). This is *entirely* due to the larger sample sizes which have the effect of reducing the standard error and thus the width of the confidence intervals.

Re-running our previous test for differences between the 1 person and 3 person household types now produces the results shown in Table 9.

In this case:

- the difference (effect size) is = 352 W or 68%;
- the 95% confidence interval for the test is now = −396.47 to −307.10 representing *much less* uncertainty/variation;
- *p* value of < 0.000 representing a *very low* risk of a false positive result as it passes all conventional thresholds.

As a result we can conclude that the large difference (or intervention effect in a trial context) is statistically significant at any conventional *p* value threshold.

Re-running our previous test for differences between the 1 person and 4+ person households now produces the results shown in Table 10.

In this case:

Table 8
Number of households and summary statistics per group (large sample).

N people	mean W	SD W	N households
1	159.22	151.75	88
2	285.12	63.89	149
3	511.00	279.36	308
4+	417.35	267.69	495

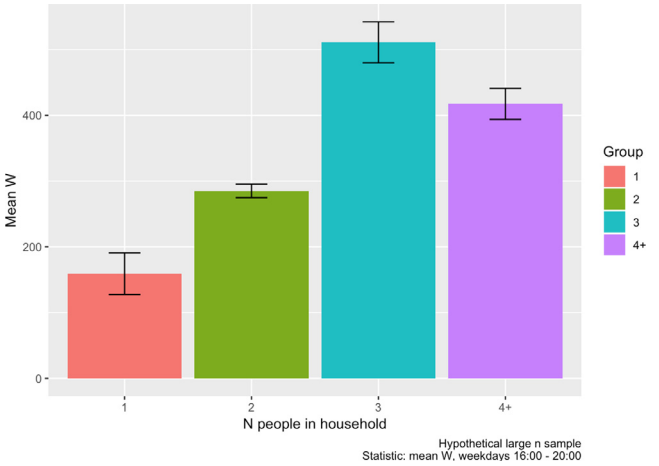

Fig. 4. Mean W demand per group for large sample (Error bars = 95% confidence intervals for the sample mean).

Table 9
T test results (1 vs 3 person households, large sample).

1 person mean	3 person mean	Mean difference	statistic	*P* value	95% CI (lower)	95% CI (upper)
159.22	511.00	–351.78	–15.50	0.000	–396.47	–307.10

- the difference (effect size) is = 258 W or 62%;
- the 95% confidence interval for the test is now = −297.89 to −218.38 representing *much less* uncertainty/variation;
- the *p* value of < 0.000 represents a *very low* risk of a false positive result as it passes all conventional thresholds.

As a result, if this had been a demand response trial, we would now conclude that 'Intervention 1' had a marginally larger effect than 'Intervention 2', the expected variation in the effect sizes is reasonably narrow (narrow confidence interval) and the risk of a Type I (false positive) error is extremely small in each case. Which option to take might then depend on other factors such as cost or ease of implementation but as we can see, sample size matters to the decisions we take...

5. Conclusion and policy implications

This paper has attempted to clarify the role of statistical significance and statistical power in the design, implementation and analysis of energy studies in general and demand response trials in particular. To do this we reported a worked example using data from the New Zealand Green Grid Household Power Demand study [3]. The results of the worked example lead us to provide the following guidance which echo arguments made elsewhere over the last 25 years [8,15,31] and in recent calls for improving the rigour of social science energy research [16,25].

5.1. Statistical power and study design

As we have shown, the choice between implementing a study from which we can conclude little with any certainty and one from which we

Table 10
T test results (1 vs 4 person households, large sample).

1 person mean	4+ person mean	Mean difference	statistic	*P* value	95% CI (lower)	95% CI (upper)
159.22	417.35	–258.13	–12.80	0.000	–297.89	–218.38

may be able to conclude something with some certainty boils down to appropriate sample sizing via statistical power analysis.

Unfortunately sample size cannot be fixed later on – all we can do is patch the leaky boat by shifting the power or statistical significance goal posts. In the interests of transparency and good research practice we therefore need to *get it right first time* and conduct statistical power analysis before we start to make sure the study is even worth trying. If we don't have previous data to use, we must *justify* our choices through power analysis based on defensible assumptions and/or decisions on the minimum effect size we need to justify an intervention [14]. This is hardly news [7,8] but it certainly seems to be news in energy studies [13,25]. We have provided a worked example that can be used to guide study designs and readers with a need are encouraged to peruse the relevant chapters in Glennerster and Takavarasha's excellent introduction [14].

Whether or not it is worth the increased cost of running a larger study depends on the precision required and the eventual decision makers' tolerance of Type I and Type II error risks. If our interventions really did result in a reduction of 40-60% then clearly a much smaller study could be used (c.f. Table 2). However effect sizes of this magnitude are virtually unknown in the energy efficiency and demand response literatures with few robust studies reporting effect sizes larger than 15% [10,12,26]. As a result the power analyses presented in Table 2 offer a realistic guide to the likely sample sizes that will be required under both conventional ($p < 0.01$ or $p < 0.05$) and more relaxed Type I error tolerances ($p < 0.1$ or $p < 0.2$) in the context of New Zealand heat pump demand at least. It remains for other studies to present results for other appliances or for whole-household demand and in other countries. We would be very surprised if the results turned out to be substantially different.

5.2. Reporting tests, making inferences and taking decisions

When making inference and subsequently taking decisions, we must therefore pay attention to all three of these elements *always*:

- average effect size: what is the *average bang for buck*?
- effect size confidence intervals: *how uncertain is the bang*?
- the *p* value: *what is the risk of a false positive*?

This means that we need to report all three elements, *always*. We should also report the statistical power assumption used so that we are clear on the risk of Type II errors. Taken together these elements enable the assessment of the substantive significance of the results. Again, we have provided worked examples that researchers and analysts can use to guide their reporting.

5.3. Best practice

Based on this guidance, professionals tasked with designing and analysing such studies should implement good sample design based on statistical power analysis. They should then provide a nuanced and fully reported analysis based on an understanding of what test statistic effect sizes, confidence intervals and *p* values can tell them.

Commercial or public policy decision makers can use this guidance to help them make evidence-based and defensible commercial strategy or policy intervention decisions based on the nuanced reporting provided by analysts. This will encourage decision-makers not to ignore practically useful results just because they do not meet normative thresholds inherited from the medical sciences and so help to avoid throwing the substantive baby out with the *p*-value bathwater.

Project managers, who are often caught between the analyst rock and the decision maker hard place can use this guidance to understand what can count as evidence, for what purpose and in what context. This knowledge can be used to effectively manage study resources and to develop a robust, contextually meaningful and *defensible* strategy for making decisions and for convincing others to make decisions. It will also help them to avoid focusing only on those results which are 'statistically significant' even if small in magnitude and of little practical use. Ideally this will enable all stakeholders to distinguish the inappropriate search for statistical significance from the need for actionable evidence.

Finally, and more importantly for the field, following this guidance will also enable future analysts to compare previous results with their own. This will ensure that all studies contribute to the development of a robust evidence base rather than adding yet another to the current scattering of isolated and incomparable results [13,16,25].

Acknowledgments

We would like to thank collaborators and partners on a number of applied research projects for prodding us into thinking about these issues more deeply and clearly than we otherwise would have done. We hope this paper helps to bring some clarity.

We would especially like to thank Greg Overton of BRANZ Ltd (New Zealand) for pointing out data processing errors that affected our initial power demand calculations.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 700386 [*SPATIALEC*] and the UK Office of Gas and Electricity Markets [under the *Solent Achieving Value from Efficiency* project].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.erss.2019.101260.

References

[1] AECOM, Energy Demand Research Project: Final Analysis, AECOM, St Albans, 2011.
[2] V. Amrhein, S. Greenland, B. McShane, Scientists rise up against statistical significance, Nature 567 (7748) (2019) 305, https://doi.org/10.1038/d41586-019-00857-9.
[3] B. Anderson, D. Eyers, R. Ford, D.G. Ocampo, R. Peniamina, J. Stephenson, K. Suomalainen, L. Wilcocks, M. Jack. 2018. New Zealand GREEN grid household electricity demand study 2014-2018, September. doi:10.5255/UKDA-SN-853334.
[4] C.F. Camerer, A. Dreber, E. Forsell, T.H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, Evaluating replicability of laboratory experiments in economics, Science 351 (6280) (2016) 1433–1436.
[5] CER, Smart Meter Electricity Consumer Behaviour Trial Data, Irish Social Science Data Archive, Dublin, 2012http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.
[6] S. Champely, 2018. *pwr: Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr.
[7] J. Cohen, Statistical Power Analysis for the Behavioral Sciences 2 Erlbaum, NJ: Hilsdale, 1988.
[8] J. Cohen, Things i have learned (so far), Am. Psychol. 45 (12) (1990) 1304.

[9] J. Cohen, Statistical power analysis, Curr. Dir. Psychol. Sci. 1 (3) (1992) 98–101, https://doi.org/10.1111/1467-8721.ep10768783.
[10] M.A. 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 (October) (2013) 729–739, https://doi.org/10.1016/j.enpol.2013.05.109.
[11] M. Dowle, A. Srinivasan, T. Short, S. Lianoglou with contributions from R. Saporta, and E. Antonyan. 2015. *Data.Table: Extension of Data.Frame*. http://CRAN.R-project.org/package=data.table.
[12] E.R. Frederiks, K. Stenner, E.V. Hobman, Household energy use: applying behavioural economics to understand consumer decision-making and behaviour, Renew. Sustain. Energy Rev. 41 (January) (2015) 1385–1394, https://doi.org/10.1016/j.rser.2014.09.026.
[13] E.R. Frederiks, K. Stenner, E.V. Hobman, M. Fischle, Evaluating energy behavior change programs using randomized controlled trials: best practice guidelines for policymakers, Energy Res. Soc. Sci. 22 (December) (2016) 147–164, https://doi.org/10.1016/j.erss.2016.08.020.
[14] R. Glennerster, K. Takavarasha, Running Randomized Evaluations: A Practical Guide, Princeton University Press, 2013.
[15] S. Greenland, S.J. Senn, K.J. Rothman, J.B. Carlin, C. Poole, S.N. Goodman, D.G. Altman, Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations, Eur. J. Epidemiol. 31 (4) (2016) 337–350, https://doi.org/10.1007/s10654-016-0149-3.
[16] G.M. Huebner, M.L. Nicolson, M.J. Fell, H. Kennard, S. Elam, C. Hanmer, C. Johnson, D. Shipworth, Are we heading towards a replicability crisis in energy efficiency research? A toolkit for improving the quality, transparency and replicability of energy efficiency impact evaluations, Proceedings of the European Council for an Energy Efficient Economy ECEEE, 2017.
[17] J.P.A. Ioannidis, M.R. Munafò, P. Fusar-Poli, B.A. Nosek, S.P. David, Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention, Trends Cognit. Sci. 18 (5) (2014) 235–241, https://doi.org/10.1016/j.tics.2014.02.010.
[18] J.PA. Ioannidis, Why science is not necessarily self-correcting, Perspect. Psychol. Sci. 7 (6) (2012) 645–654.
[19] H.C. Kraemer, C. Blasey, How Many Subjects?: Statistical Power Analysis in Research, Sage Publications, 2015.
[20] S.E. Maxwell, The persistence of underpowered studies in psychological research: causes, consequences, and remedies, Psychol. Methods 9 (2) (2004) 147.
[21] H. Pashler, E.J. Wagenmakers, 'Editors' introduction to the special section on replicability in psychological science: a crisis of confidence?', Perspect. Psychol. Sci. 7 (6) (2012) 528–530, https://doi.org/10.1177/1745691612465253.
[22] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015http://www.R-project.org/.
[23] Rocky Mountain Institute. 2006. 'Automated Demand Response System Pilot: Final Report'. https://www.smartgrid.gov/files/Aumated_Demd_Response_System_Pilot_Volume_1_Intro_Exec_Summa.pdf.
[24] J. Schofield, Dynamic Time-of-Use Electricity Pricing for Residential Demand Response: Design and Analysis of the Low Carbon London Smart-Metering Trial, Imperial College London, London, 2015.
[25] B.K. Sovacool, J. Axsen, S. Sorrell, Promoting novelty, rigor, and style in energy social science: towards codes of practice for appropriate methods and research design, Energy Res. Soc. Sci. (2018), https://doi.org/10.1016/j.erss.2018.07.007 October.
[26] A. Srivastava, S. Van Passel, E. Laes, Assessing the success of electricity demand response programs: a meta-analysis, Energy Res. Soc. Sci. 40 (June) (2018) 110–117, https://doi.org/10.1016/j.erss.2017.12.005.
[27] P.E. Tressoldi, D. Giofré, F. Sella, G. Cumming, High impact= high statistical standards? Not necessarily so, PloS One 8 (2) (2013) e56180.
[28] UKPN, The Final Energy Saving Trial Report, UK Power Networks, London, 2017http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Energywise/.
[29] UKPN, The Energy Shifting Trial Report, UK Power Networks, London, 2018http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Energywise/.
[30] V. White, M. Jones, Warm, Dry, Healthy? Insights from the 2015 House Condition Survey on Insulation, Ventilation, Heating and Mould in New Zealand Houses'. SR 372. BRANZ Study Report, BRANZ Ltd., Porirua, New Zealand, 2017.
[31] R.L. Wasserstein, N.A. Lazar, The ASA's statement on p-values: context, process, and purpose, Am. Stat. 70 (2) (2016) 129–133, https://doi.org/10.1080/00031305.2016.1154108.
[32] H. Wickham, Ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2009http://ggplot2.org.
[33] H. Wickham, R. Francois. 2016. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.
[34] Y. Xie, 2016. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.