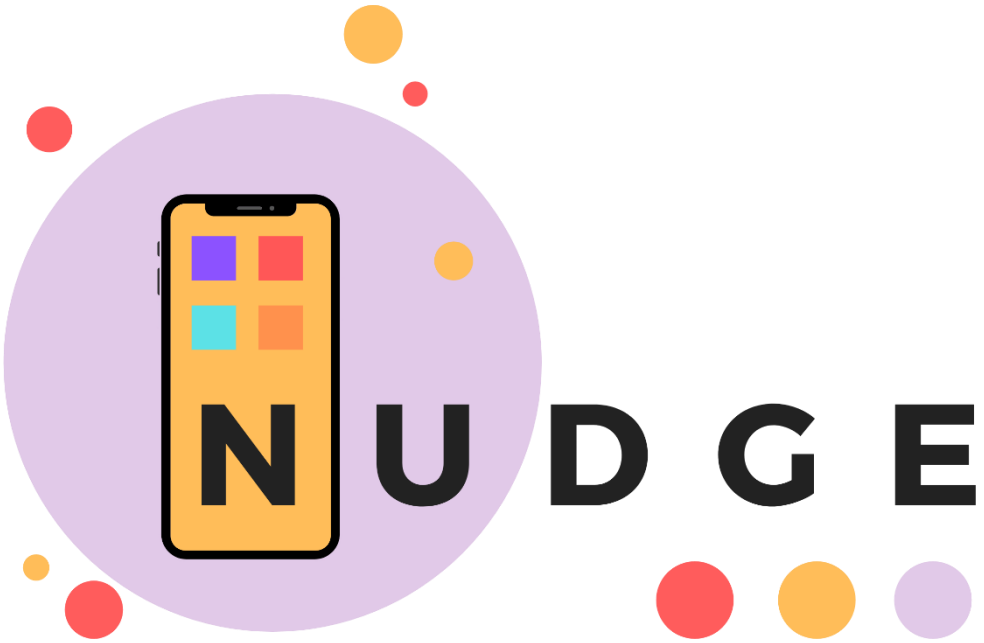


Nudging consumers towards energy efficiency through behavioural science

Research methodology for assessing the effectiveness of interventions regarding change of energy efficient behavior (D2.2)

Deliverable Information

Nature: Public

Version: 1 (Month 8)

Delivery date: 05/07/2021

Project Coordinator: Filippos Anagnostopoulos, IEECP, filippos@ieecp.org

Authors: Stephanie Van Hove (Stephanie.Vanhove@UGent.be), Peter Conradie (Peter.Conradie@UGent.be), Merkouris Karaliopoulos(mkaralio@aueb.gr), Sabine Pelka (sabine.pelka@isi.fraunhofer.de)

Project information

Project Title	Nudging consumers towards energy efficiency through behavioural science
Project Acronym	NUDGE
Project Number	927012
Project dates	September 2020 – August 2023

Rev.	Written by	Date	Checked by	Date
1	Stephanie Van Hove (IMEC), Peter Conradie (IMEC), Merkouris Karaliopoulos (AUEB-RC), Sabine Pelka (Fraunhofer)	01/07/ 2021	Merkouris Karaliopoulos (AUEB-RC) Werner Neumeier (beegy) Polychronis Symeonidis (DOMX)	05/07 /2021

Legal Notice

About

Efforts to induce energy-friendly behaviour from end-users through behavioural interventions are characterized by a lack of customer personalization ("one-size-fits-all interventions"), a partial understanding about how different interventions interact with each other and contrasting evidence about their effectiveness, as a result of poor testing under real world conditions.

NUDGE has been conceived to unleash the potential of behavioural interventions for long-lasting energy efficiency behaviour changes, paving the way to the generalized use of such interventions as a worthy addition to the policy-making toolbox. We take a mixed approach to the consumer analysis and intervention design with tasks combining surveys and field trials. Firmly rooted in behavioural science methods, we will study individual psychological and contextual variables underlying consumers' behaviour to tailor the design of behavioural interventions for them, with a clear bias towards interventions of the nudging type.

The designed interventions are compared against traditional ones in field trials (pilots) in five different EU states, exhibiting striking diversity in terms of innovative energy usage scenarios (e.g., PV production for EV charging, DR for natural gas), demographic and socio-economic variables of the involved populations, mediation platforms for operationalizing the intervention (smart mobile apps, dashboards, web portals, educational material and intergenerational learning practices).

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957012.

Project partners

Contents

Project information 2
Legal Notice 2
About 3
Project partners 3
1. Introduction 5
2. Timeline 5
3. General research protocol and background 5
3.1 Study design: Randomized Control Trials 5
3.1.1 Control group 5
3.1.2 Three intervention designs 6
3.2 Sample sizes and recruitment of participants 9
3.2.1 Sample size 9
3.2.2 Inclusion and exclusion criteria for participant recruitment and assignment to groups 11
4. Experimentation phases 12
4.1.1 Pre-intervention phase (or baseline) 12
4.1.2 Treatment phase 12
4.1.3 In-between treatment period of within-subjects experiments 12
5. Design of Experiments 13
5.1 Overview 13
5.2 Spring-Stof 15
5.3 ZEZ, INEGI and domX 16
5.4 Beegy 18
6. Measuring instruments 19
7. Analytic Strategy 20
Bibliography 21
Appendix I 22

1. Introduction

This document introduces the proposed methodology for assessing the effectiveness of the interventions. We present and discuss several methodological approaches, ranging from randomised controlled trials to A/B testing. This includes a brief (theoretical) discussion about different methodological approaches, with accompanying power calculations.

Most importantly, we present the five pilots with general guidelines, most significantly because the details of the interventions might still change. In addition to the experimental setup, we will also formulate the first research hypothesis per pilot and provide an overview of possible outcome variables per pilot.

2. Timeline

The timing for this task is June 2021 to August 2023. All five pilots in NUDGE are planned to start by M9 and their execution follows an identical three-phase time plan:

- **Pre-interventions phase**: This phase will enable the establishment of benchmarks (baselines) in terms of energy use and consumer behaviour of the participating households in each pilot. This initial phase will last 5 months (M11-M14).
- **Testing phase**: This second phase includes the actual testing of the planned interventions in each pilot. It will have a prolonged duration, between M15 and M32, so as to provide for the execution of several (often consecutive) interventions and their evaluation. However, while this task runs until M32 (April 2023), we expect data gathering to be finalised by March 2023, to allow ample time for data processing and analysis.
- **Post-interventions phase**: In this last phase, planned for the time interval between M33 and M36, the pilots will keep on running in the absence of any interventions. The aim is to evaluate whether the consumers maintain improved energy-efficient behaviour after nudges cease, thus gaining insights into the long-lasting impact of the NUDGE approach.

3. General research protocol and background

3.1 Study design: Randomized Control Trials

A randomized controlled trial (RCT) is a 'rigorous, scientific experiment purposely designed to test the efficacy of an intervention on a sample of participants drawn from some target population' [1]. As discussed by [1] RTCs are optimal because they allow assessment of whether cause and effect exist between treatment and outcome, while also *assessing the validity, utility and overall cost-effectiveness of an intervention, relative to business-as-usual or alternative interventions.* RCTs have the following characteristics:

- Participants are **randomly assigned** to two distinct and non-overlapping groups, called the *treatment* and the *control* group, to warrant the high internal validity of the experiment [1].
- Only households that are willing to participate are part of the randomization. Therefore, households that do not wish to participate cannot be considered the control group. This is important to avoid self-selection bias.

3.1.1 Control group

A control group consists of people who do not receive the treatment. Participants of an experiment are ideally randomly assigned to either the treatment or control group, or matched on relevant criteria, e.g., baseline consumption in the context of energy consumption. The inclusion of a control

group allows to isolate the dependent variable, e.g., energy conservation performance. Any differences between the treatment and control group are then caused by the manipulation of the independent variable, e.g., the introduction of nudges. From a statistical point of view, researchers evaluate if the result found in the treatment group is significantly different (or not) from the result in the control group (mostly with a confidence level of p = .05, or the probability of 5% that a given result is completely random and it has not been induced by the experiment).

Ideally, the control group should be subject to 'business-as-usual' [2]. In the context of the NUDGE interventions this could be, for instance, the use of electricity meters or an electronic platform without nudges. This means that the control group should only differ in the parameter to be evaluated: the intervention at hand. Hence, a minimum version of the electronic platform (evidently, without nudges) should be provided in the control group. Otherwise, this only results in knowing whether using an electronic platform (with nudges) is better (or worse) than not using a platform at all.

3.1.2 Three intervention designs

3.1.2.1 Between-subjects design

In a between-subjects design, the total sample is subdivided, either into random or matched groups (whereby it is attempted to keep both groups characteristically similar). These are some typical characteristics of the between-subjects design:

- This design is also typically known as 'pretest-posttest control group design' [3]. The inclusion of a control group allows for (1) the evaluation of behavioural change related to the mere lapse of time; and, (2) the comparison of attitudinal outcomes [2].
- The between-subjects design also limits learning effects because the groups and interventions are independent. However, a significant downside of between-subjects design is the need for larger sample sizes.
- The design lets exploring the **single vs. additive vs. interactive effects** of combining treatments [1]. For example, interactive effects are at play when the original effect of nudge 1 is enhanced (or mitigated) when combined with nudge 2. Additive effects can be determined if the single effect of nudge 1 remains unchanged when combined with nudge 2. Testing nudges in isolation and in combination is the only way to compare the relative effectiveness of the individual nudges and thus determine the most impactful nudge of the intervention.

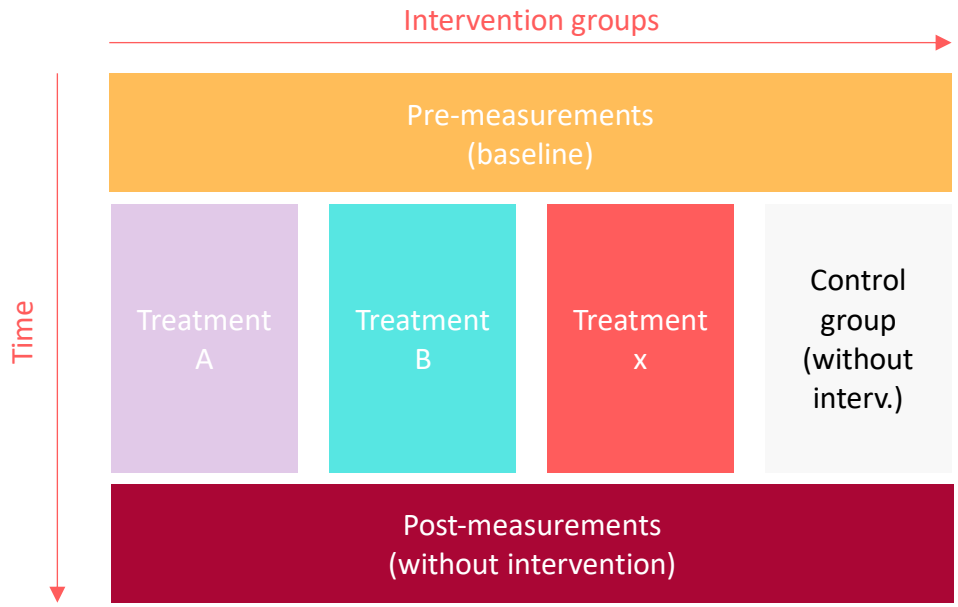


Figure 1: Illustration of a between-subjects design that compares different treatments between different groups of participants. All groups fill out a similar pre- and post-measurement (yellow and burgundy blocks). With regard to the intervention the overall sample is divided into k groups, with all but one group being subject to different treatments (purple to red block). One group serves as the control group (grey block).

3.1.2.2 Within-subjects design

Within-subjects designs don't employ a traditional control group, but typically compare the pre- and post-interventions phase. Balanced designs, by contrast, introduce half the population with an intervention and half not, with the groups swapped around. These are some issues and characteristics of within-subjects designs:

- **Confounding effects**: They cannot eliminate the possibility that the retrieved impact is due to external factors, such as weather conditions, increased energy pricing, or a pandemic period.
- **Multiple-treatment interference**: Participants are sequentially exposed to more than one treatment, which makes it impossible to identify the precise effect of each single treatment [1]. This means that the combination of interventions may have 'additive, interactive, or even counteractive effects' [1], or the effect could largely be attributed to one aspect.
- **Combined long-term effects**: Only the combined effect of nudges can be evaluated in the long term, since the treatment group has not been exposed to a particular nudge in isolation. This renders it impossible to determine the relative importance of the nudges.
- **Short-term effect 'in isolation'**: After having administered nudge A, a treatment-free period is included to examine the short-term effects immediately after the treatment and after several weeks. Thereafter, nudge B is administered. Please note for multiple-treatment interference after a period.

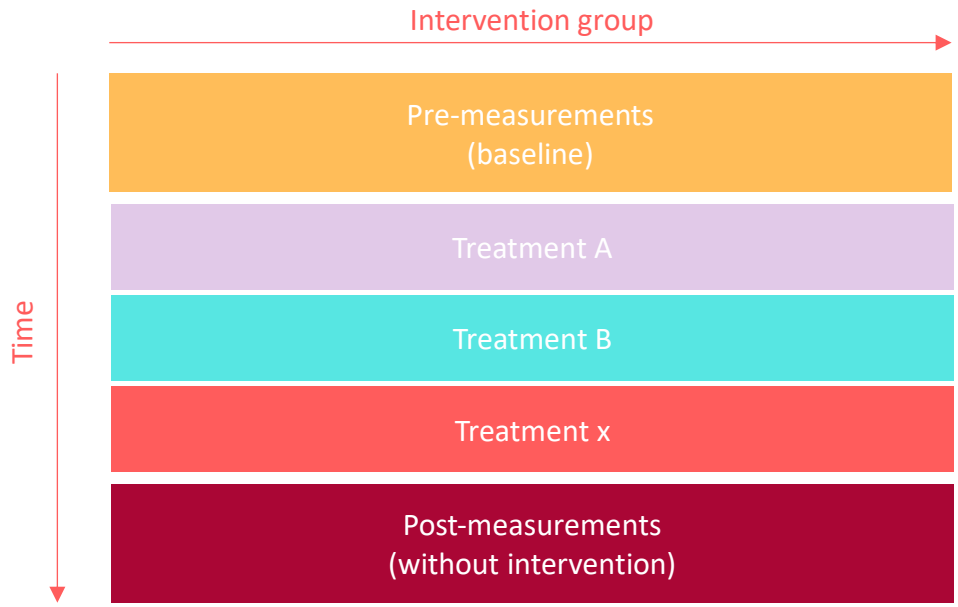


Figure 2: Illustration of a within-subjects design that compares different treatments within the same group of participants. All groups fill out a similar pre- and post-measurement (yellow and burgundy blocks). Regarding the intervention (purple to red block), all participants take part in all interventions at different times. No control group is included in a pure within-subjects design, as participants serve as their own control by providing baseline scores (see yellow block).

3.1.2.3 Mixed-subjects design

Mixed-subjects design combine elements from both within- and between-subjects designs. This allows assessing changes in consumption for individuals (i.e., through assessment of longitudinal repeated measurements) while also allowing comparisons between a treatment and a control group. Mixed-subjects designs, thus, share most of the drawbacks and benefits of between- and within-subjects designs.

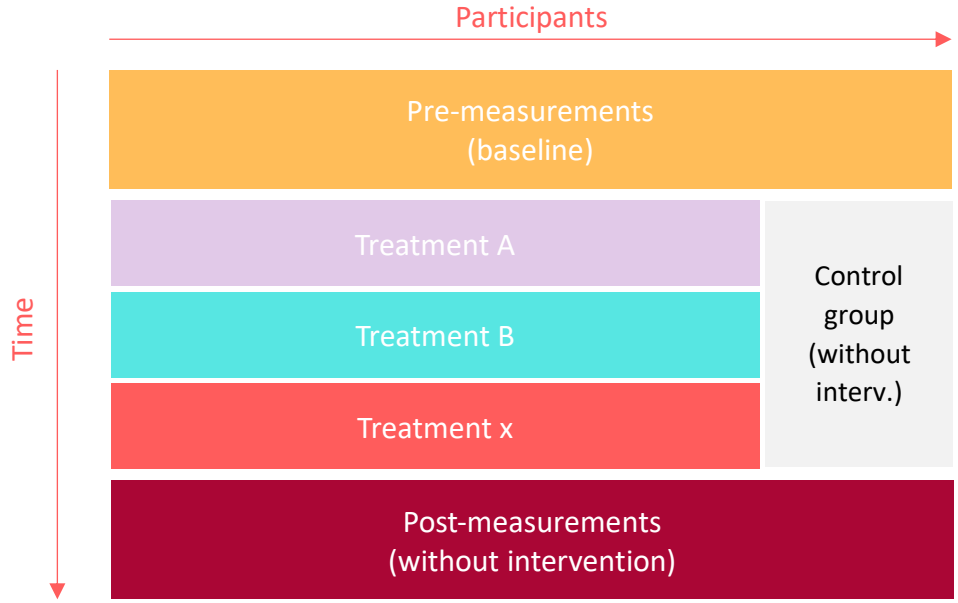


Figure 3: Illustration of a mixed-subjects design. All groups fill out a similar pre- and post-measurement (yellow and burgundy blocks). Regarding the intervention, one group of participants takes part in all interventions at different times (purple to red block). Next to the treatment group, also a control group is included (different from a within-subjects design, grey block).

3.2 Sample sizes and recruitment of participants

3.2.1 Sample size

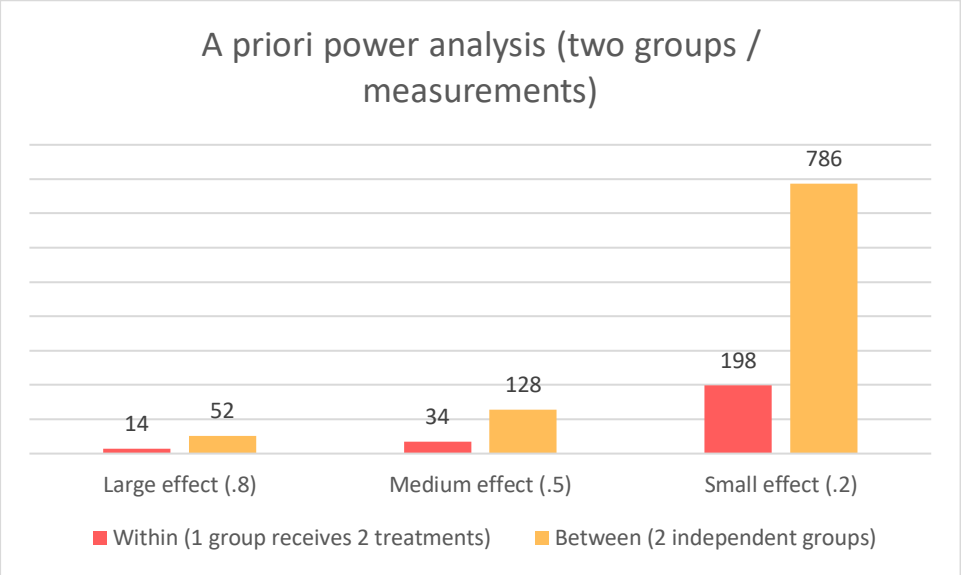


Figure 4: A priori power analysis given alpha is set to .05, power to .80 and 2 conditions with effect sizes (Cohen's d) ranging large to small

The rule of thumb of 20 households/participants per condition is considered an absolute minimum, however, more advanced statistical analysis requires 30 households/participants per condition [2]. Moreover, the required sample sizes depend on the expected effect sizes and can be determined through *a priori* power analysis. When fixing explanatory power (to 0.8, see [4, p. 54] for an expanded discussion about the .8 power threshold), statistical significance α (to α = 0.05) and treatment number k (to k = 2) constant and varying expected effect sizes from large (f = 0.8) to small (f = 0.2), the required sample sizes for a within-subjects design (paired-samples t-test, see **Error! Reference source not found.**) range from 14 to approximately 200 participants, whereas for a between-subjects design (t-test, see **Error! Reference source not found.**) they range from 26 to approximately 400 participants per group. More information on power analyses can be found in [5]. We would recommend a total sample size of 20 households *times the number of nudges* as the absolute minimum and 30 households times the number of nudges as the desired sample size.

Table 1: Example effect sizes of previous intervention studies executed in Europe within the domain of energy conservation.

Previous RCTs	Number of groups	Control	Total sample size	Effect size	Energy consumption
Crago (2020)	3	yes	62	$f^2 = 0.13$	+14.2%
Kendel (2017	2	Yes	65	$f^2 = 0.75$	-13 à -23.3%
Delmas (2014)	3	Yes	66	$f^2 = 0.12$	-20%
Tiefenbeck (2016)	3	yes	636	$f^2 = 0.59$	-5% (energy) -22% (water)
Kandul (2020)	5	yes	821	$f^2 = 0.01$	-1.2% (indoor temperature)
…					

Likewise, for a repeated measures ANOVA study (i.e., within-subjects design that compares different treatments applied at different times to the same group of participants), we find that sample size

requirements become far smaller for similar effect sizes, when compared to the one-way analysis of variance for three groups.

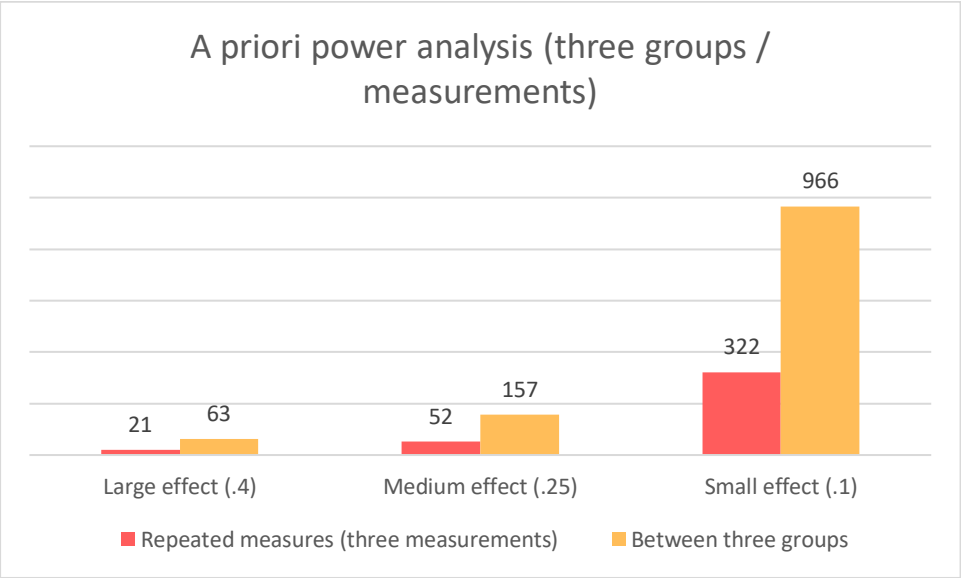


Figure 5: A priori power analysis given alpha is set to .05, power to .80 and 3 conditions with effect sizes (Cohen's f) ranging from large to small

Looking more specifically at the available samples within the pilots (*N* = ≤ 100 participants), we can perform power calculations if we assume a small effect size (f = .01). Furthermore, we reduce groups from 3 to 2, increasing participants per group from 33 to 50. The results, shown in **Error! Reference source not found.** illustrate the statistical power for both variations to be .13 and .17, respectively. These numbers fall short of the .8 threshold we consider acceptable for power (and which is proposed by Cohen, 2013). Figure 6 visualizes this analysis, using groups of 33 and 50 participants, respectively.

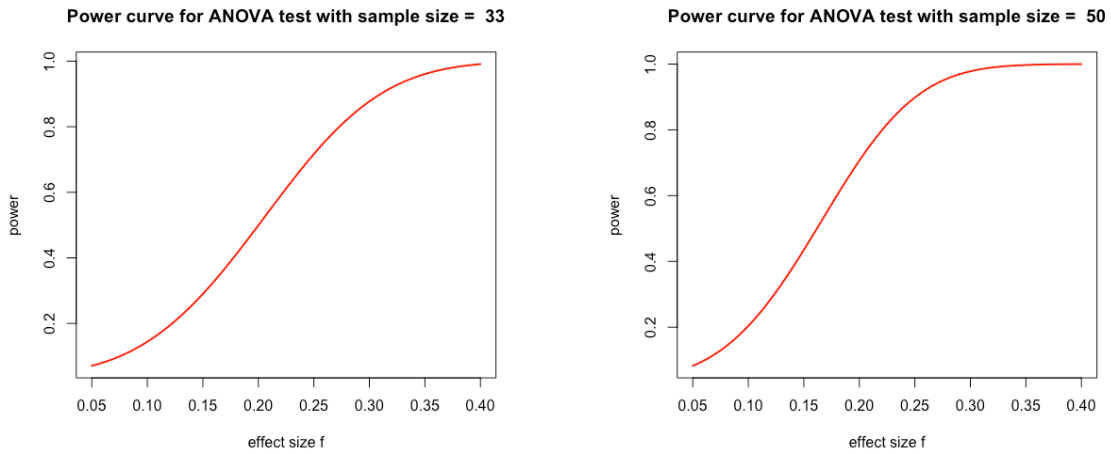


Figure 6: Power curve, assuming a group size of 33 and 50 respectively

Alternatively, we can set the statistical power at .8, the recommended level proposed by Cohen (2013), and keep sample sizes the same at either 33 or 50 per group. This analysis, also shown in Table 2, illustrates that this experimental design demands effect sizes of .32 and .28 respectively, which can be classified as medium to large. This is also visible in Figure 6, which relates all three quantities (sample size, efffect size and power) together.

Table 2: Power analysis of experiments (row 1-2: assuming small effect size f = .01; row 3-4: assuming recommended power of .8)

Number of groups	Participants per group	Total sample size	Effect size	Power
3	33	~100	.1	.13
2	50	100	.1	.17
3	33	~100	.32	.8
2	50	100	.28	.8

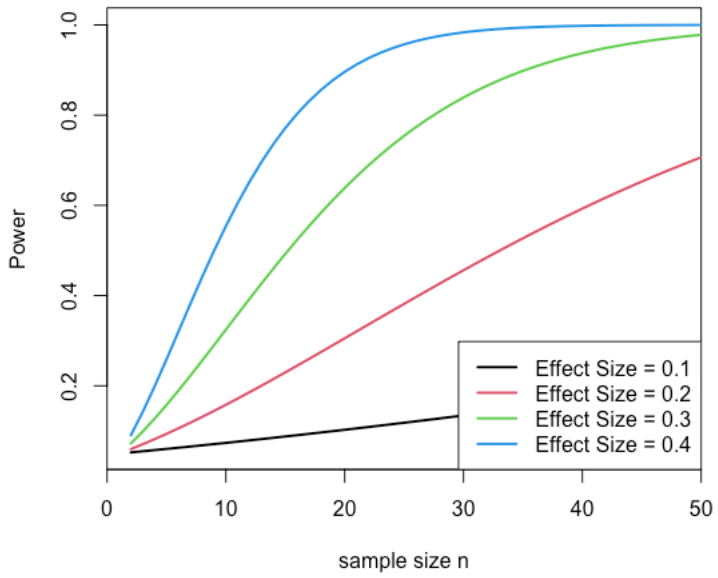


Figure 7: Power curve, assuming effects ranging from small (.1) to large (.4)

In summary, these results for different types of experimental designs suggest that with the current sample size limitations and power set at .8, we can find statistically reliable results only with effects sizes around .3.

3.2.2 Inclusion and exclusion criteria for participant recruitment and assignment to groups

The pilots should strive for a diverse sample of households that reflects the relevant population of the study. In the case of NUDGE, a relevant population might correspond to all energy consumers within a particular geographical region. Parameters that may differ across participants involve:

- Energy consumption: low versus high energy consumption
- Family situation: with or without children? single people? number of family members?
- Living area: small and large houses?
- Housing type: apartment, row house, (semi-)detached house
- House property: owned/rented residence
- House heating: presence of thermostat, manual/programmable/smart thermostat

3.2.2.1 Inclusion criteria

- General: head of household should at least be 18 years old
- Portuguese pilot:
    - Family with young children (children aged 0-12 years at the beginning of the pilot)
    - Living in the district of Porto

    - Using electricity as main energy vector
    - Having Wi-Fi at home
- Inclusion criteria yet to be defined for the other pilots (Belgium, Germany, Croatia, and Greece)

3.2.2.2 Exclusion criteria

- Non-ratepaying households, e.g., households who pay for their use of heating energy and thus have no individual meter (in case of multi-apartment buildings)
- Non-family household, e.g., group of students living together

3.2.2.3 Assignment of participants to conditions

Randomization of participants to keep groups as coherent as possible in terms of household type, age (of the head of household), and living area.

4. Experimentation phases

4.1.1 Pre-intervention phase (or baseline)

A pre-intervention phase (in combination with a post-intervention phase, including data logging processes) is an absolute requirement for three reasons: (1) to control for differences between the experimental and control group at the onset of the intervention, (2) to evaluate the impact of the nudges on energy-related behaviour and attitudes, and (3) to control for characteristics of drop-outs (i.e., attrition analysis).

No rule of thumb exists for determining the ideal duration to measure baseline consumption, the intervention period, and in-between periods of treatments. In order to provide some insight in practices of previous studies, we have conducted an exploratory analysis of RCTs that provided feedback on a weekly, daily or real-time basis (*n* = 18, subset of studies reviewed in T1.1). Other studies that have evaluated feedback on a monthly basis (e.g., Home Energy Reports) or single efforts such as letters or flyers have not been considered since these studies are inherently different in their research design and typically last longer.

This analysis shows that baseline periods vary from 1 week to 52 weeks with 57% of studies implementing a **baseline period** of 4 weeks or less (MED = 3.0, M = 12.3, SD = 18.1). The one-week baseline study implemented an in-home display during three months and evaluated the intervention after one week and three months (Schultz et al., 2015). On the other end of the 'baseline duration' continuum, Loock and colleagues evaluated in 2011 and 2013 a web portal with a baseline of 52 weeks, and an intervention duration of respectively 6 (no control group) and 20 weeks (inclusion of control group) (Note that the collection of one year of energy use data prior to the intervention, in the context of energy consumption, is recommended in [1] in order to be able to control for seasonal effects). Both baseline extrema found significant conservation effects.

4.1.2 Treatment phase

The **intervention duration** varies from 1 week [6] to 35 weeks [7] with 33% of studies implementing a treatment period of 4 weeks or less (MED = 7.0, M = 10.4, SD = 8.9).

4.1.3 In-between treatment period of within-subjects experiments

Only one study implemented a within-subjects design (see [8]) and consequently, reports an **in-between treatment period** of 4 weeks.

Based on these insights, we recommend the following minimum requirements regarding the duration of each of the experimental phases:

Pre-intervention phase (during M11-M14):

- Baseline (pre-intervention) duration: 3 months

Testing phase (during M15-M32):

- Treatment period per group: 4 weeks
- In-between treatment period: 3 weeks (with regard to consecutive interventions, this in-between period can serve as the baseline for the next intervention of the pilot)

5. Design of Experiments

5.1 Overview

Given the heterogeneous nature of the five pilots, we partition the experiments under the five pilots in three groups, proposing re-usable building blocks that can be used for each mini-experiment or treatment. The first group involves the Belgian pilot, run by Spring-Stof (see section 5.2) Due to the different status of their sample (i.e., its educational context), they will apply a hybrid experiment with a control group that will fill out a pre- and post-measurement. The second group comprises the Portuguese, Greek and Croatian pilots and is amenable to a within-subjects design (see section 5.3). For the last group, corresponding to the German pilot, we suggest a between-subjects design (where possible) due to its technical limitations, such as the inability of removing nudges once implemented, trade-off between feasibility and costs (see section 5.4). However, given tec

Table 3: Detailed overview of the needs and requirements per pilot

T2.2 - Questions regarding the interventions					
Numbers (#)	domX	ZEZ	Springstof	INEGI	BEEGY
Number of treatments, i.e. (combinations of) nudges	±5	Approx. 5 - 8 (depending on the technical complexity of the APP that can be covered within ZEZ's budget in the Project)		From 5 to 7, depending on what can be done with the apps that will be developed.	As mentioned, we are planning for 3-4 phases. In each phase several nudges will be introduced, differentiated by the two tools, I.e., a web portal (non EV-owners) and charging app (EV-owners).
What is the total size of your pilot? Include also people who might not be provided with a technological intervention, but where usage might still be measured.	100	100	50	100	100, divided into 2 groups (50 each)
Of the total group, how many will receive a smart meter / application / intervention?	100	100 smart meters; all will have access to the APP; from the beginning of testing, 50 households will form the non-intervention group, and other 50 households will be in the intervention group	50 (all)	100 with smart meters; all will have access to the APP; two protocols are possible: (A) either from the beginning of testing 50 households will be devided into non-intervention group, and other 50 households will be in the intervention group and will	All 100 will receive the Webportal. In addition, 50 will also receive the charging app.

				have also an IEQ multi-sensor system; or, (B) hHaving all 100 participants receiving nudges.	
Must have (yes/no)					
At the end of the study, will every participant or participant group be exposed to all the nudges in your pilot? (sequential exposure)	yes	no	yes	If protocol (A) is chosen, NO. If (B), YES.	yes. The only difference will be, that only 50 will have access to the charging app.
Will there be a control group who do not receive any nudges at all?	no	yes	no	If (A) YES, if (B) NO.	no (we rather plan for a pre-intervention phase, in which all participants are using the tools without any nudges
Are you able to remove nudges once they are introduced?	yes	no	no	Yes, if app allows.	Not yet clear. We have that as a requirement. However, this depends on the assessment of technical feasibility and cost, which is currently evaluated
Can you remove all nudges from your pilot and revert to the "baseline" software version?	yes	no		Yes, if app allows.	with the current system: no
Can you introduce multiple nudges to your pilot simultaneously? (simultaneous exposure)	yes	yes	yes	Yes	yes, this is the plan
Might have (yes/no)					
Besides the control group, is it possible to divide the sample in different groups who receive distinct nudges? (isolated exposure)	yes	no	no	Yes	no yet clear. We have that as a requirement. However, this depends on the assessment of technical feasibility and cost, which is currently evaluated

Moreover, since pilots differ substantially in the number of experiments, we recommend a modular approach for the second group of pilots, where the total nudge experiment consists out of several mini-experiments, varying only in sample makeup (i.e., samples are newly randomised from scratch after each trial), intervention and duration. However, we emphasise the modularity of our experimental design, allowing the German pilot to also use the within-subjects design.

Finally, the majority of our experiments rely on energy consumption as outcome variable, i.e., we hypothesise that a nudge will reduce energy consumption. Moreover, other external parameters that have a significant impact on the result of the experiment, such as outside temperature, can be

controlled for when inserted in a formula with small sample sizes (or through a covariate in regression analysis with large sample sizes).

A significant confounding variable throughout all our experiments will be the weather. For example, sunny days might lead to less time spent indoor, with commensurate decreases in energy consumption. Likewise, extremely hot days might result in more time spent indoors with air conditioning, increasing consumption. While within subjects designs control for personal variability (i.e.: sensitivity to a particular nudge), the impact of the weather will remain. However, using an analytic approach such as multilevel modelling (described in section 7) we can eliminate random factors such as the weather. Moreover, for each mini-experiment (see Figure 8) data can be analysed separately as if it were a between subject design and thus controlling for the weather.

5.2 Spring-Stof

For the Spring-Stof experiment a between-subjects design will be implemented with paired data from pupils and parents.

- **Group 1 – Control group (n=50).** This group will consist of parents whose children will not be subjected to the course. The control group will only fill out a pre- and post-test. A pre-test is needed to ensure that the control and treatment group are equivalent in their composition. A post-test is needed to evaluate if there is a significant difference in energy knowledge among parents compared to the treatment group.
- **Group 2 – Treatment (n=50)**. This second group will consist of parents and their children, with the latter receiving education about energy consumption at school. Both parents and children will fill out a pre- and post-test.

Estimated timeline:

- **T0 –** Control (parents) and treatment group (parents and pupils) both fill out a pre-test questionnaire, i.e., the pilot-specific questionnaire, focused on awareness and knowledge concerning energy consumption. Hereafter, the course on energy consumption and conservation takes off (+/- 5 sessions during 6 months). Data collection among the treatment groups on energy consumption starts.
- **Tx –** Control (parents) and treatment group (parents and pupils) both complete the second questionnaire that consists of the outcome variables.

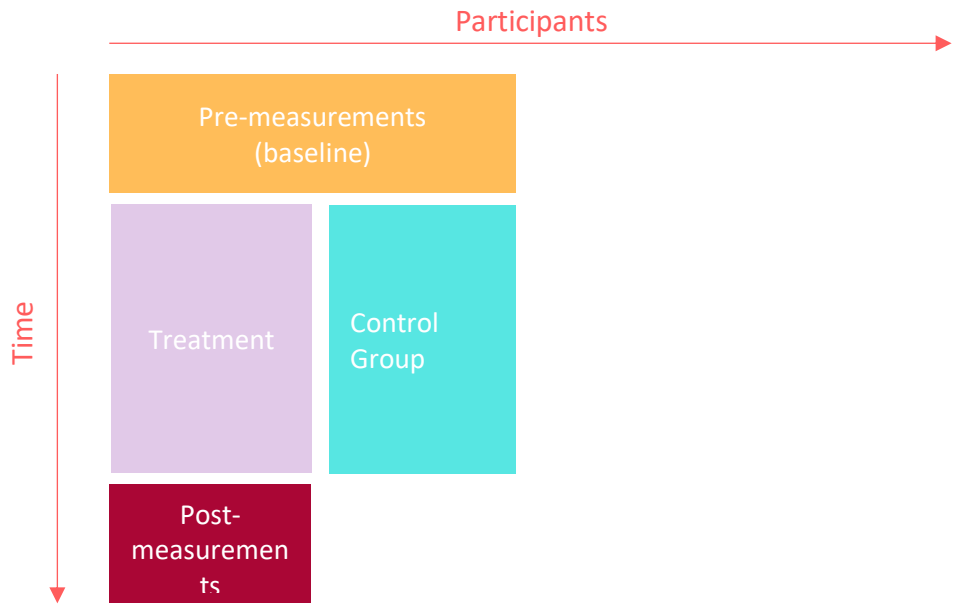
Outcome variables: the outcome variables will be evaluated in the pre-test and post-test. Both tests will be filled out by parents in the control and treatment group, and by pupils in the treatment group. Through the provision of paired data, i.e., from pupils and parents, we are able to evaluate the impact from intergenerational education (see [9]) on self-reported measures such as, awareness and knowledge of energy consumption, and logging of energy consumption from gas and electricity. Awareness largely focuses on people's awareness of their own energy consumption [10]. Energy knowledge covers the factual aspect of people's energy literacy [10].

Hypotheses:

H1. Energy conservation education for pupils positively impacts awareness of energy consumption in (a) pupils and (b) their parents.

H2. Energy conservation education for pupils positively impacts energy-related knowledge in (a) pupils and (b) their parents.

H3. Energy conservation education for pupils positively impacts energy conservation behaviour in households.



5.3 ZEZ, INEGI and domX

As noted above, we recommend a within-subjects research design for the Greek, Croatian and Portuguese pilots. The figure below illustrates this setup. First, two random groups are created, k_1 (n = 50) and k_2 (n = 50). First, k_1 is exposed to the nudge, while k_2 acts as a control group. Following this, after a predetermined treatment (i.e., four weeks), groups get swapped with k_2 being exposed to the nudge, and k_1 becoming the control group. Each experiment evaluates one nudge and lasts 8 consecutive weeks. After the experiment has terminated, the households are randomized again and assigned to one of the two groups of the following experiment that starts after three weeks. Concretely, this means that a maximum of 7 interventions can take place during the testing phase (M15-M32) if pilots adhere to the recommended durations (see 4. Experimentation phases).

Estimated timeline of testing phase (during M15-M32):

	Intervention 1								In-between treatment period			Intervention 2								
w	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Ti
k1	T	T	T	T	C	C	C	C				T	T	T	T	C	C	C	C	..
k2	C	C	C	C	T	T	T	T				C	C	C	C	T	T	T	T	..

Outcome variables:

The impact of the nudges on participant's energy-related behaviour and perception will be evaluated by means of three data sources:

1. Sensor data: objective measurement that measures final energy consumption;
2. App data: objective measurement that serves as a proxy of awareness and acceptance of the intervention;
3. Survey data: self-reported data to evaluate perceived behavioural change. Administered before and after each intervention, and before and after the complete trial.

Ideally, sensor, app, and survey data are triangulated to contextualize behavioural change.

Greek pilot:

- Energy consumption from gas (Wh)

Croatian pilot:

- Energy consumption and production from electricity (Wh)

Portuguese pilot:

- Air quality parameters: indoor temperature, indoor relative humidity, carbon dioxide (CO2), particulate matter ($PM_{2.5}$, PM_{10}), volatile organic compounds (VOCs) (see
- [11])
- Perceived indoor environmental quality (self-reported, see [11] and [12] for a review of studies)
- Energy consumption from electricity (Wh)

Hypotheses:

Greek pilot:

- H1. The energy conservation nudges decrease energy consumption in households.
- H2. The energy conservation nudges increase the energy consumption efficiency.

Croatian pilot:

- H1. The energy conservation nudges decrease energy consumption in households.
- H2. The energy conservation nudges increase the self-consumption share in households having PV panels.

Portuguese pilot:

- H1. The energy conservation nudges decrease energy consumption in households.
- H2. The energy conservation nudges positively impact indoor environmental quality.
- H3. Changes in indoor environmental quality improves both the participants' perception of air quality in their homes and the compliance of the indoor levels with existing guidelines
    - H3a. Changes in indoor environmental quality improves the participants' perception of air quality in their homes.
    - H3b. Changes in indoor environmental quality improves the compliance of indoor levels with existing guidelines.

–

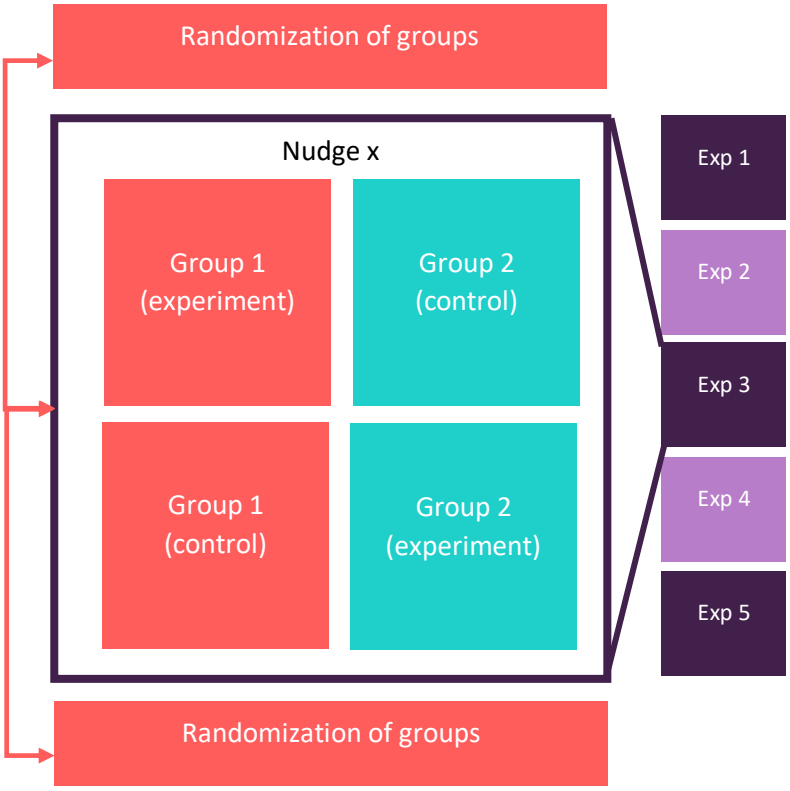


Figure 8: Illustration of the within-subjects design proposed for the ZEZ, INEGI and domX pilots

5.4 Beegy

Tentatively at the time of writing, the German pilot will apply a layered approach to their application development, whereby features are introduced one after the other. Once a feature is implemented, it cannot (easily) be removed. As a result, a within-subject design is not feasible, because participants cannot be placed into the control group after exposure. Given this limitation, we foresee two possible scenarios for Beegy.

In **Scenario 1**, (see [8] for an application) we capture baseline data prior to the introduction of the first nudge. For each successive nudge, data is captured and analysed using a repeated-measures approach. While this is a within-subjects design, it lacks counterbalancing and is thus less robust than a research design that repeatedly subjects the same participants to treatments (as applied with ZEZ, ENEGI and domX). We subsequently also lack a control group. Therefore, it is not possible to establish whether external factors (i.e., confounding factors, due to weather conditions, a pandemic situation, ...) will impact the results.

In **Scenario 2**, we apply a more classic A/B design, with half of the sample receiving successively the update (i.e., treatment). We retain a control group throughout the study, so we can evaluate if the result found in the treatment group is significantly different (or not) from the result in the control group. Additionally, we propose subject matching, whereby the resulting groups have roughly the same characteristics (i.e., baseline consumption, attitude towards energy saving, demographics, etc).

Hypotheses:

- H1. The energy conservation nudges increase the self-consumption share in households having PV panels.

    - H1a. Households owning an electric vehicle increase the self-consumption share by optimizing charging strategies.
    - H1b. Households that have no electric vehicle increase the self-consumption share by load shifting usage of white goods.
- H2. The energy conservation nudges *decrease the perceived effort* of load shifting in households having PV panels.
- H3. The energy conservation nudges *increase the motivation* to consider load shifting in households having PV panels.
- H4. The energy conservation nudges *increase awareness* of the importance of load shifting in households having PV panels.

6. Measuring instruments

Measurements at all locations will happen through application of sensors, with slight deviations (i.e., some will only measure electricity, others only gas, or a combination). INEGI will additionally measure a variety of other information, including carbon monoxide (CO), carbon dioxide (CO2). More specifically, the measurements to take place in each pilot include:

Belgian pilot:

- Energy consumption from gas and electricity (Wh)
  (data logging through smart meter or weekly/monthly meter readings)
- Awareness of energy consumption (self-reported measurement, [10])
- Energy-related knowledge
  (self-reported measurement [10], [13])
- Where possible, usage data of the smart meter and or smartphone applications.

Greek pilot:

- Energy consumption from gas (Wh)
- Where possible, usage data of the smart meter and/or smartphone applications.

Croatian pilot:

- Energy consumption from electricity (Wh)
- PV energy production (Wh)
- Where possible, usage data of the smart meter and/or smartphone applications.

Portuguese pilot:

- Air quality parameters: indoor temperature, indoor relative humidity, CO2, $PM_{2.5}$, PM_{10} and VOCs (see [11])
- Perceived indoor environmental quality (self-reported, see [11], [12] for a review of studies).
- Energy consumption from electricity (Wh)
- Where possible, usage data of the smart meter and/or smartphone applications.

German pilot:

- Energy consumption from electricity and EV (Wh)
- PV energy production (Wh)
- Perceived effort level, perceived usefulness of information, perceived fit-to-daily routine, perceived motivation level (self-reported measurements)

– Where possible, usage data of the smart meter and or smartphone applications.

7. Analytic Strategy

The inclusion of a control group and/or baseline also determines the data analysis approach. Some studies report conservation effects with reference to baseline consumption, whereas others compare treatment group performance to control group performance. An exploratory analysis of RCTs (*n* = 18, subset of studies reviewed in T1.1) demonstrates that approaches largely fall apart in three categories:

1. Comparison reference = control group
    - In combination with controlling for baseline consumption [14][15]
2. Comparison reference = baseline measurement.
    - No control group included: [16][14]
    - In combination with controlling for consumption in control group. This can either be done with an equation (see [17][18]) or as covariate.
3. Comparison with both control group and baseline measurement [19] .

Given these complexities, we favour multilevel models, linear mixed models, mixed effects models, or hierarchical linear models. This statistical approach can be used for the analysis both of nested data (i.e.: math results of children in different classes) or longitudinal, repeated measures data [20], [21], as is the case in NUDGE. Especially for within-subject designs, we violate the assumption of sample independence, i.e., participants in both measurement groups are related, or in fact the same. Given this, Ordinary Least Squares (OLS) analysis are not appropriate. By contrast, by using a multilevel model, we can include our participant as random effect in our analysis. Furthermore, extraneous variables like temperature can additionally be included as random effect.

Bibliography

[1] E. R. Frederiks, K. Stenner, E. V. Hobman, and M. Fischle, "Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers," *Energy Res. Soc. Sci.*, vol. 22, pp. 147–164, 2016, doi: 10.1016/j.erss.2016.08.020.

[2] A. All, "Digital Game-based Learning Under the Microscope: Development of a Procedure for Assessing the Effectiveness of Educational Games Aimed at Cognitive Learning Outcomes," Ghent University, 2016.

[3] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*, vol. 107. Houghton Mifflin Company, 1963.

[4] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic Press, 2013.

[5] A. Hunt, "A Researcher's Guide to Power Analysis," vol. 0. Utah State University, pp. 1–11, 2012.

[6] K. Ito, T. Ida, and M. Tanaka, "Moral Suasion and economic incentives: Field experimental evidence from energy demand," *Am. Econ. J. Econ. Policy*, vol. 10, no. 1, pp. 240–267, 2018, doi: 10.1257/pol.20160093.

[7] A. Kendel, N. Lazaric, and K. Maréchal, "What do people 'learn by looking' at direct feedback on their energy consumption? Results of a field study in Southern France," *Energy Policy*, vol. 108, no. June, pp. 593–605, 2017, doi: 10.1016/j.enpol.2017.06.020.

[8] C. L. Crago, J. M. Spraggon, and E. Hunter, "Motivating non-ratepaying households with feedback and social nudges: A cautionary tale," *Energy Policy*, vol. 145, no. July, p. 111764, Oct. 2020, doi: 10.1016/j.enpol.2020.111764.

[9] P. Damerell, C. Howe, and E. J. Milner-Gulland, "Child-orientated environmental education influences adult knowledge and household behaviour," *Environ. Res. Lett.*, vol. 8, no. 1, p. 015016, Mar. 2013, doi: 10.1088/1748-9326/8/1/015016.

[10] B. Stikvoort, P. Juslin, and C. Bartusch, "Good things come in small packages: is there a common set of motivators for energy behaviour?," *Energy Effic.*, vol. 11, no. 7, pp. 1599–1615, Oct. 2018, doi: 10.1007/s12053-017-9537-0.

[11] J. Wilson *et al.*, "Watts-to-Wellbeing: does residential energy conservation improve health?," *Energy Effic.*, vol. 7, no. 1, pp. 151–160, Feb. 2014, doi: 10.1007/s12053-013-9216-8.

[12] M. A. Ortiz, S. R. Kurvers, and P. M. Bluyssen, "A review of comfort, health, and energy use: Understanding daily energy use and wellbeing for the development of a new approach to study comfort," *Energy Build.*, vol. 152, pp. 323–335, Oct. 2017, doi: 10.1016/j.enbuild.2017.07.060.

[13] B. Sütterlin, T. A. Brunner, and M. Siegrist, "Who puts the most energy into energy conservation? A segmentation of energy consumers based on energy-related behavioral characteristics," *Energy Policy*, vol. 39, no. 12, pp. 8137–8152, 2011, doi: 10.1016/j.enpol.2011.10.008.

[14] P. W. Schultz, M. Estrada, J. Schmitt, R. Sokoloski, and N. Silva-Send, "Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms," *Energy*, vol. 90, pp. 351–358, Oct. 2015, doi: 10.1016/j.energy.2015.06.130.

[15] E. Myers and M. Souza, "Social comparison nudges without monetary incentives: Evidence from home energy reports," *J. Environ. Econ. Manage.*, vol. 101, p. 102315, May 2020, doi:

10.1016/j.jeem.2020.102315.

[16] J. E. Petersen, V. Shunturov, K. Janda, G. Platt, and K. Weinberger, “Dormitory residents reduce electricity consumption when exposed to real-time visual feedback and incentives,” no. January, 2007, doi: 10.1108/14676370710717562.

[17] G. Peschiera and J. E. Taylor, “The impact of peer network position on electricity consumption in building occupant networks utilizing energy feedback systems,” *Energy Build.*, vol. 49, pp. 584–590, 2012, doi: 10.1016/j.enbuild.2012.03.011.

[18] V. Tiefenbeck, T. Staake, K. Roth, and O. Sachs, “For better or for worse ? Empirical evidence of moral licensing in a behavioral energy conservation campaign,” *Energy Policy*, vol. 57, pp. 160–171, 2013, doi: 10.1016/j.enpol.2013.01.021.

[19] G. Peschiera, J. E. Taylor, and J. A. Siegel, “Response – relapse patterns of building occupant electricity consumption following exposure to personal , contextualized and occupant peer network utilization data,” *Energy Build.*, vol. 42, no. 8, pp. 1329–1336, 2010, doi: 10.1016/j.enbuild.2010.03.001.

[20] R. Tarling, *Statistical Modelling for Social Researchers: Principles and Practice*. 2009.

[21] G. Verbeke, G. Molenberghs, and D. Rizopoulos, “Random Effects Models for Longitudinal Data,” in *Longitudinal Research with Latent Variables*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 37–96.

Appendix I

```
library("pwr")

#code for paired and two sample t-tests

pwr.t.test(d = 0.8, power = 0.80, sig.level = 0.05, type = c("paired"))

pwr.t.test(d = 0.8, power = 0.80, sig.level = 0.05, type = c("two.sample"))

pwr.t.test(d = 0.5, power = 0.80, sig.level = 0.05, type = c("paired"))

pwr.t.test(d = 0.5, power = 0.80, sig.level = 0.05, type = c("two.sample"))

pwr.t.test(d = 0.2, power = 0.80, sig.level = 0.05, type = c("paired"))

pwr.t.test(d = 0.2, power = 0.80, sig.level = 0.05, type = c("two.sample"))


library("WebPower")

# code for one way anova

pwr.anova.test(k=3,f=.1,sig.level=.05,power=.8)

pwr.anova.test(k=3,f=.25,sig.level=.05,power=.8)

pwr.anova.test(k=3,f=.4,sig.level=.05,power=.8)


# code for repeated measured anova

wp.rmanova(ng=1, nm=3, f=0.4, nscor=1, alpha=.05, power=.8,type=1)

wp.rmanova(ng=1, nm=3, f=0.25, nscor=1, alpha=.05, power=.8,type=1)

wp.rmanova(ng=1, nm=3, f=0.1, nscor=1, alpha=.05, power=.8,type=1)
```