Pre-Analysis Plan
‘Hot Water Challenge’

Alexander Goetz[1], Harald Mayr[1,*], and Renate Schubert[1]

[1]ETH Zurich
[*]Corresponding author, harald.mayr@econ.gess.ethz.ch

November 7, 2019

Abstract

Various interventions have been shown to effectively reduce private households' energy consumption. We evaluate a randomized controlled trial in the context of the 'Hot Water Challenge', an intervention focused on hot water consumption, to address three research questions. Can regret enhance the effectiveness of traditional lotteries as a tool to incentivize households? For which type of household are social comparison interventions effective? Does an intervention focused on hot water consumption lead to spillovers on cold water or heating energy consumption?

Keywords: RCT, hot water consumption, regret lottery, social comparison, spillovers

Contents

Chapter 1

The experiment

This chapter describes aspects of the experiment that are common among the three research questions in subsequent chapters 2, 3, and 4. The order of these chapters does not imply a priority ranking.

1.1 Collaborating parties

1.1.1 Owning Company

The experiment is conducted in buildings owned by one of the largest real estate owners in Switzerland, called 'Owning Company' for the purposes of this document. The Owning Company collaborates with the research team to pursue its sustainability goals and see these efforts covered by the media. While the Owning Company had a crucial role in enabling the experiment, it does not actively manage the buildings and the relationship with renters. Hence, its involvement during the actual implementation of the experiment is limited. Subjects learn about the role of the Owning Company in the first email.

1.1.2 Managing Company

Real estate management is performed by the 'Managing Company'. The Managing Company is responsible for the relationship with tenants, including advertising of vacant objects, property management, and all rental paperwork. The subjects of the experiment see the Managing Company as responsible for the 'Hot Water Challenge', the sender of all emails, and their point of contact for anything regarding the interventions.

1.1.3 Metering Company

The Managing Company mandates the 'Metering Company' to measure utilities consumption. The Metering Company maintains appropriate metering systems for the (yearly) accounting of service charges. The main task of the Metering Company during the 'Hot Water Challenge' is to provide monthly hot water consumption data for all households in our sample early in the subsequent month. Because the data is read remotely, the Metering Company does not become visible towards the subjects of the experiment.

1.1.4 Implementing Company

The email campaign is implemented by a specialized agency, the 'Implementing Company'. The Implementing Company implements the randomization, prepares all mass mailings, collects data about these mailings, and serves as a partner for technical aspects of the intervention. Subjects do not learn about the involvement of the Implementing Company.

1.1.5 Research team

The research team coordinates and supports all involved parties to implement the experiment. Subjects do not learn about the involvement of the research team until the 'Hot Water Challenge' has ended.

1.2 Experimental design

The field experiment addresses three distinct research questions in a cross-randomized design. The *Control* group (20% of the sample[1] does not receive any intervention. The remaining 80% of the sample receive individual hot water consumption feedback, a 5% savings target, and intervention components randomized along two dimensions:

- The type of lottery (*StandardLottery* vs. *RegretLottery*, described in detail in chapter 2)
- Social comparison (*NoComparison* vs. *Comparison*, described in chapter 3)

The effects of these aggregated interventions, as compared to the *Control* group, are subject of chapter 4.

The research design is summarized in table 1.1. The five groups (A, B, C, D, *Control*) are of equal size, i.e. 20% of the sample.

[1] The relative size of the *Control* group reflects the Managing Company's preference for large treatment groups. It yields comparable power for realistic treatment effects of the different research questions (see section 1.5).

1.2.1 Subject pool

Our sample of 4,775 households consists of private households living in buildings owned by the Owning Company who

- have per-household metering systems that deliver monthly remote readings for hot water consumption,
- had valid hot water readings for August 2019,
- had an email address in the database of the Managing Company before the start of the intervention,
- did not terminate their rental agreement before the start of the intervention,
- and rented exactly one object before the start of the intervention.

1.2.2 Intervention emails

Each household in one of the intervention groups (A, B, C, D) receives a total of 10 emails:

- Email *basic information*: this email was the same for each group, describing the intent and nature of the 'Hot Water Challenge'. The intervention is framed as an initiative by the Owning Company, carried out by the Managing Company. The email contains links to view legal details on the lottery, information on data privacy, hot water saving recommendations, and to opt out of future emails regarding the 'Hot Water Challenge'.
- Email *baseline consumption*: this email shows the household's hot water consumption in the baseline month September 2019.
- Emails *consumption* 1-4: these emails illustrate the consumption of hot water in the months October 2019 to January 2020, as well as the consumption history since September 2019. They also inform the household whether the respective month's savings target was reached.
- Emails *draw* 1-4: these emails arrive one day after the household learns whether it reached last month's savings target. They convey information about the monthly lottery draw and the continuation (Email *draw* 1-3) or conclusion (Email *draw* 4) of the intervention.

	StandardLottery	*RegretLottery*	
NoComparison	A	B	
Comparison	C	D	
			Control

Table 1.1: Experimental design

The exact details of the emails' content varies by intervention group, as described in sections 2.2 and 3.2. The graphical illustrations of hot water consumption data are created by the Implementing Company, using a Python script prepared by the research team.

1.2.3 Institutional review board

The experiment was approved by the Ethics Commission of ETH Zurich on July 22, 2019.

1.3 Data

1.3.1 Consumption data

The Metering Company described in subsection 1.1.3 delivers monthly hot water consumption data. The data for a given month are scheduled to be delivered by day 9 of the subsequent month or the first working day thereafter. Hot water consumption of household i in a given month is denoted by HW_i^{month}.

Cold water consumption CW_i^{month} and heating energy consumption HE_i^{month} will be delivered after the intervention ends. For technical reasons, we expect that HE_i^{month} will only be available for a subsample.

1.3.2 Household characteristics

The following household characteristics are available before the intervention phase starts:

- $Rooms$: number of rooms (in steps of 0.5, censored at 6)
- $RentCategory$: monthly net rent (CHF) in categories $< 1,000$; $1,000$ to $1,499$; $1,500$ to $1,999$; $2,000$ to $2,499$; $2,500$ to $2,999$; $3,000$ to $3,499$; from $3,500$
- $Female$: gender of the main tenant
- $AgeCategory$ of the main tenant: < 30; 30 to 39; 40 to 49; 50 to 59; 60 to 69; from 70
- $Language$: German; French; Italian
- $BuildingID$: identifies the building
- $BuildingSize$: the number of observed households in the building

1.3.3 Email data

As the intervention proceeds, detailed information about the email campaign become available for each email sent:

- *Bounced*: binary indicator for bounced email (i.e. email not delivered)
- *Opened*: binary indicator for downloaded remote email content
- $ViewedLotteryDetails$: binary indicator for legal lottery details link opened
- $ViewedDataPrivacy$: binary indicator for data privacy link opened
- $ViewedRecommendation$: binary indicator for hot water saving recommendations link opened
- *OptOut*: binary indicator for opt-out link opened

1.4 Randomization

The research team provided a custom-made Python script for stratified randomization, but randomization was actually *implemented* by the Implementing Company on September 20, 2019. As described in section 1.6, the research team will remain blind to the treatments until December 2019.

Randomization was performed within strata of the following household characteristics:

- $Rooms >= 3$ (binary)
- $Rent >= CHF\ 1,500$ (binary)
- $Age >= 40$ (binary + unknown)
- *Language* (German; French; Italian)
- *BuildingSize* above median (binary)
- $HEmetered$ (binary, whether heating consumption is metered)
- HW^{aug} (in quintiles)

Randomization was performed in two steps to accommodate two levels of randomization:

1. The *Control* group was randomly allocated on the building level.
2. All other groups were randomly allocated on the household level.

This approach was chosen to prevent conceived unfairness if some households in a building would participate in the lottery while others in the same building could not. Stratified randomization on the mentioned variables was performed in both steps.

1.5 Power simulation

Power simulation allows us to obtain more credible power estimates than traditional power calculations, incorporating unique features of the distribution beyond mean and variance (e.g. non-normal data, outliers) as well as the stratified randomization strategy.

1.5.1 Setup

Power simulations were conducted with hot water consumption data for July and August 2019. $1,000$ randomization procedures were simulated according to subsection 1.4.

For each of these $1,000$ group allocations, the August consumption of each household in one subsample (the 'treated' group) is reduced by a certain percentage (i.e. the simulated treatment effect). We then calculate power as the share of group allocations for which the percent change in hot water consumption from July to August 2019 is different in a two-sided test with significance level 0.05.

We calculated $1,000$ p-values for the Mann-Whitney-U-test and the t-test.

1.5.2 Outlier exclusion

An important finding of the power analysis is that outliers in the data have important implications for power, especially for the t-test. Excluding outliers generally boosts power of the t-test for reasonable exclusion procedures. Modest exclusion of outliers also improves power of the Mann-Whitney-U-test somewhat. The following approach yields good power with reasonable restrictions:

1. Excluding the lowest 5% hot water consumption values in the baseline period (these cases often lead to enormous percent changes).

2. Excluding observations with the lowest and highest 1% in the percent change from baseline (after the exclusion in 1).

We will employ this sample restriction for the entire analysis described in this document. We will use the Mann-Whitney-U-test for our main hypothesis tests. Our rationale for choosing the Mann-Whitney-U-test over the t-test is described in section 2.5.

1.5.3 *StandardLottery* vs. *RegretLottery*, *Comparison* vs. *NoComparison*

This subsection concerns the power simulations for chapters 2 and 3. The unit of observation for the corresponding tests is the household. The test *StandardLottery* vs. *RegretLottery* is equivalent to *Comparison* vs. *NoComparison* because all four treatment groups are of equal size (see section 1.2).

A hypothetical 3.15% reduction in hot water consumption can be detected with 80.3% (Mann-Whitney-U-test) and 59% (t-test) power.

The main outcome in chapters 2 and 3 covers four months instead of the one month we consider in this power simulation. We expect to be sufficiently powered for substantially smaller effects, as some noise should cancel out over the four months of data.

1.5.4 *Control* vs. Treatments (i.e. all other groups)

This subsection concerns the power simulations for chapter 4. Because the *Control* group is randomized at the building level (see subsection 1.4 for details), the unit of observation for the corresponding tests is the building level.

A hypothetical 4.8% reduction in hot water consumption can be detected in 79.8% (Mann-Whitney-U-test) and 67.1% (t-test).

The main outcome in chapter 4 covers four months instead of the one month we consider in this power simulation. We expect to be sufficiently powered for substantially smaller effects, as some noise should cancel out over the four months of data.

1.6 Timeline

The research team has no access to personal or pseudonymized data at any time. All data for the research team will be fully anonymized to make sure that individual households cannot be identified. The research team will remain blind to the treatments until December 2019.

Figure 1.1 shows the timeline with respect to measurement, including the sending of each scheduled email. The following data deliveries are scheduled:

1. Hot water consumption data for July and August 2019 merged with household characteristics (already took place on August 12 and September 10, 2019),

2. Hot water consumption data updates for the following months in the respective subsequent month (scheduled around day 9, depending on weekends),

3. Unblinding of the intervention groups (scheduled with the data delivery in December),

4. Cold water and heating consumption data for the months September 2019 to January 2020 in February 2020.

This pre-analysis plan is archived well before the intervention groups are unblinded to the research team and before the relevant consumption data become available. We archive it at `socialscienceregistry.org`, the American Economic Association's registry for randomized controlled trials, on November 7, 2019.

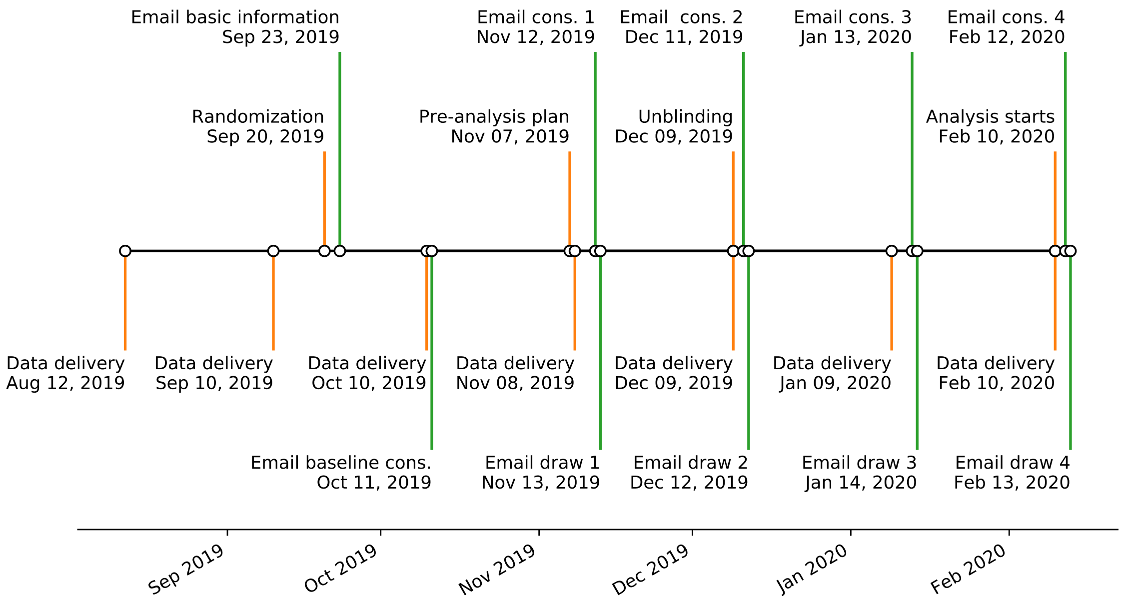

Figure 1.1: Timeline

Chapter 2

Regret for lotteries

This section describes the first of three research questions in the field experiment.

2.1 Research question and hypotheses

Lotteries are widely used to incentivize behavior (Lange et al., 2007; Landry et al., 2006; Evans et al., 1988). The success of lotteries can be explained by a well-known behavioral bias: people tend to overweight small probabilities (Steiner and Stewart, 2016; Snowberg and Wolfers, 2010).[1] This chapter's research question asks whether another behavioral peculiarity — regret aversion — can make lotteries even more effective. In a recent study, van der Swaluw et al. (2016) show that participants in a lottery with anticipated regret (see below for an explanation of this concept) attained their attendance goals more often than participants in a control group without incentive. We investigate whether regret can enhance the effect of standard lotteries.

In the standard implementation of lottery incentives, participation in the lottery is conditional on some behavior (in our case: reaching the 5% savings target). Importantly, the subject is usually only informed about the lottery outcome if she actually wins the prize. This corresponds to the first of the following four possible scenarios:

1. Reached savings target, was drawn in lottery
2. Reached savings target, was not drawn in lottery
3. Did not reach savings target, would have been drawn in lottery
4. Did not reach savings target, would not have been drawn in lottery

[1] Put simply, most people prefer a 0.05% chance to win a thousand dollars over 50 cents.

A subject can distinguish scenarios 1 and 2 if she knows that she reached the savings target. If she did not, however, scenarios 3 and 4 remain indistinguishable.

Scenario 3 is of particular interest for this research question because mere anticipation of regret may cause behavior change. Ample evidence demonstrates that humans are averse to regret (Brewer et al., 2016) and a simple twist may unleash this hidden potential in standard lotteries: providing subjects with information about the lottery outcome independent of their behavior. We hypothesize:

1. Providing subjects with information about the lottery outcome *independent of subject's behavior* (as compared to *only if the subject reached the savings target and was drawn in the lottery*) leads to a reduction in hot water consumption.

2. These treatment effects can be predicted by household characteristics available at the start of the intervention.

2.2 The treatment

In both *StandardLottery* (treatment groups A and C) and *RegretLottery* (treatment groups B and D), each household receives 10 emails in total (see subsection 1.2.2). The emails' content varies depending on the household's treatment group, the attainment of the savings target, and the outcome of the lottery draw:

Email *basic information*: The first email is sent in the month before the scheduled start of the study (September 23, 2019). It provides basic information about the 'Hot Water Challenge'. The content of the email is exactly the same for each household.

Email *baseline consumption*: This email is sent in early October after the data delivery for the baseline month September. It announces the start of the 'Hot Water Challenge' and provides the households with information on their baseline consumption. The content varies with regard to *StandardLottery* vs. *RegretLottery*. Households receive the following lottery explanation:

> "Next month we will inform you whether you have achieved your savings target. On the subsequent day, we will inform you in a separate email ...

- *RegretLottery*: ... whether you were drawn or not."
- *StandardLottery*: ... in case you were drawn and achieved your savings target."

Furthermore and again, the households have the possibility to access lottery conditions, privacy policy, and opt-out form.

Emails *consumption* 1-4: These emails are sent after each intervention month, i.e. in early November, December, January, and February. Each of these emails informs whether the household met its savings target in the preceding month and provides diagrams to illustrate the household's progress (see section 3.2 for details on the diagrams).

Emails *draw* 1-4: These emails are sent one day after each Email *consumption*. They report on the draw depending on the household's treatment group, the attainment of the savings target, and the outcome of the lottery draw::

- *RegretLottery* if drawn: 'Congratulations, you were drawn! If you achieved your savings target you will get a monthly rent.'
- *RegretLottery* if not drawn: 'This time you were not drawn in the lottery.'
- *StandardLottery* if drawn and target attained: 'Congratulations, you were drawn and win! Due to achieving your savings target you will receive a monthly rent.'
- *StandardLottery* if not drawn: No information on the draw.

Independent of lottery draw and target attainment, each Email *draw* informs how the 'Hot Water Challenge' proceeds.

2.3 Estimation sample

The estimation sample consists of the $3,814$ households not in the *Control* group except

1. households whose emails bounced (i.e. could not be delivered),
2. households opting out between Email *basic information* and Email *baseline consumption*,
3. households moving out at any point before Email *draw* 4,
4. households with any missing hot water consumption data between August 2019 to January 2020,
5. households with outlier values as specified in subsection 1.5.2.

2.4 Outcome variables

All outcome variables are based on monthly hot water consumption readings HW_i^{month} where i indicates households. The percent change between the baseline period (September) and a later month is defined as:

$$HWchange_i^{month} = \frac{HW_i^{month}}{HW_i^{baseline}} - 1 \tag{2.1}$$

The **primary outcome variable** is defined as the average percent change in hot water consumption between the baseline period (September) and the intervention periods October to January:

$$HWchange_i^{intervention} = \frac{HWchange_i^{oct} + HWchange_i^{nov} + HWchange_i^{dec} + HWchange_i^{jan}}{4} \tag{2.2}$$

$HWtarget^{month}$ indicates whether the consumption savings target was achieved in a given month:

$$HWtarget_i^{month} = \mathbb{1}_{HWchange_i^{month} < -0.05} \tag{2.3}$$

The number of target achievements serves as a **secondary outcome variable**:

$$HWtarget_i^{intervention} = HWtarget_i^{oct} + HWtarget_i^{nov} + HWtarget_i^{dec} + HWtarget_i^{jan} \tag{2.4}$$

2.5 Test of the main hypothesis

Hypothesis 1 in section 2.1 is assessed by comparing treatment groups[2] A and C (*StandardLottery*) with B and D (*RegretLottery*). The sample is defined in subsection 2.3. The main outcome is defined in subsection 2.4.

Mean values of the main outcome variable will be reported for the two groups *StandardLottery* and *RegretLottery*. Statistical significance of the difference will be assessed with the Mann-Whitney-U-test (two-sided alternative). The U-statistic and p-value of the test will be reported. Our choice for the Mann-Whitney-U-test is rationalized below.

Pitman (1949) shows that the asymptotic relative efficiency (ARE) of the Mann-Whitney-U-test relative to the t-test (i.e. the natural parametric alternative) can be arbitrarily large. The actual ARE of course depends on the distribution of our outcome variable. In our data, power simulations consistently showed the Mann-Whitney-U-test's superior power as compared to the t-test, even after excluding outliers. Moreover, the Mann-Whitney-U-test is a 'safe bet' as the theoretical minimum of the aforementioned ARE is 0.864 (Hodges

[2]See table 1.1 for reference.

and Lehmann, 1956).[3]

With balance on observable characteristics ensured by the stratified randomization procedure described in subsection 1.4, multivariate analysis is not necessary. Standard OLS would likely suffer from the same problems as the t-test.

2.6 Test of treatment effect heterogeneity

This section describes the test of hypothesis 2 in section 2.1. Conditional average treatment effects (CATEs) will be estimated using causal forests (Wager and Athey, 2018; Athey et al., 2019) as implemented in the R package `grf`. We proceed as follows:

1. Replace missing values in *AgeCategory* with the variable's mode (all other predictor variables have no missing values).

2. Randomly split the sample in two subsamples of equal size, S_{train} and S_{test}, with `random.seed=0`. The split procedure is stratified on a set of household characteristics based on the list in subsection 1.4: The quintiles of HW_{aug} are replaced by household level indicators for above-median sum and above-median variance of HW_{aug} and HW_{sep}; *Gender* is added to the list.

3. Train a causal forest on S_{train}, using the `grf` function `causal_forest` with the following arguments

    - `num.trees=1000000`
    - `honesty=TRUE`
    - `tune.parameters=TRUE`
    - `num.fit.trees=10000`
    - `num.fit.reps=1000`
    - `num.optimize.reps=10000`
    - `seed=0`

    The set of predictors includes the household level sum and variance of HW_{aug} and HW_{sep}; and the list in subsection 1.3.2 except *BuildingID*. The treatment indicator is described in section 2.2 and the outcome variable is the primary outcome variable defined in section 2.4.

4. Predict CATEs for each household in S_{test} using the model trained in 3.

[3]See Rojo and Lehmann (2012, p. 335) for details on the asymptotic theory. See section 1.5 for details on our power calculations.

5. Report appropriate statistics and plots for the predicted CATEs in S_{test}

6. Regress (OLS) the outcome variable as described in section 2.4 on the predicted CATEs, the indicator variable for *RegretLottery*, and the interaction of these two variables. A positive and significant coefficient on the interaction will be evidence for predictable treatment effect heterogeneity.

7. Given that 6 shows meaningful treatment effect heterogeneity, appropriate statistics and plots will be reported to evaluate the relationship between individual household characteristics and the predicted CATEs.

2.7 Effects over time

Effect sizes may vary over time, e.g. if subjects learn about the possibility of regret in *RegretLottery*. This possibility if assessed in two ways:

1. The analysis in section 2.5 is repeated for each monthly outcome $HWchange_i^{oct}$, $HWchange_i^{nov}$, $HWchange_i^{dec}$, $HWchange_i^{jan}$.

2. The δ coefficients in the following OLS specification provide a formal test for differences in treatment effects over time :

$$
\begin{aligned}
HWchange^{month} &= \alpha + \beta \cdot \mathbb{1}_{RegretLottery} + \gamma_{nov} \cdot \mathbb{1}_{nov} + \gamma_{dec} \cdot \mathbb{1}_{dec} + \gamma_{jan} \cdot \mathbb{1}_{jan} \\
&+ \delta_{nov} \cdot \mathbb{1}_{(nov \wedge RegretLottery)} \\
&+ \delta_{dec} \cdot \mathbb{1}_{(dec \wedge RegretLottery)} \\
&+ \delta_{jan} \cdot \mathbb{1}_{(jan \wedge RegretLottery)} + \varepsilon
\end{aligned}
\tag{2.5}
$$

2.8 Exploratory analyses

In addition to the analyses specified above, exploratory analyses may be performed without pre-registration.

Chapter 3

Social comparison

3.1 Research question and hypotheses

Social psychologists have long argued that social comparisons are a central attribute of human life. As one of the first, Festinger (1954) has shown in a multitude of studies how social comparisons inuence judgments (i.e. evaluation of self and others) and even actions. For instance, individuals assess their own health status as better if they compare themselves to others with poorer health status (Suls et al., 1991) or smokers are more likely to quit if they are connected to other smokers who have similar intentions and are on the verge to quit (Gerrard et al., 2005).

Research on individual environmental behavior provides empirical evidence on the effects of social comparison interventions. In a large-scale field experiment, Allcott and Rogers (2014) feature a neighbor comparison diagram, which compares a household's recent energy use to that of 100 neighbors with similar house characteristics. They find that large short-term energy savings fade away relatively fast after treatment ends. Brook (2011) shows that people who were informed about their above-average ecological footprint changed their subsequent behavior towards more pro-environmental choices. On the other hand and rather skeptically, Costa and Kahn (2013) argue that a social comparison feedback could backlash among certain population segments.

This project focuses on social comparison feedback, informing households about their own hot water consumption, the average consumption of households in the same room category, and the average consumption of the 20% with the lowest hot water consumption values. We hypothesize:

1. Providing subjects with with information on hot water consumption of itself *and* comparable households (as compared to information about the households own consumption *only*) leads to a reduction in hot

water consumption.

2. These treatment effects can be predicted by household characteristics available at the start of the intervention.

3.2 The treatment

Starting from the general setup described in section 2.2, we provide different consumption information in Emails *consumption* 1-4 in the treatment groups *Comparison* and *NoComparison*. *NoComparison* households receive consumption diagrams about their own consumption only. *Comparison* households receive the same diagrams augmented with measures of comparable households. Figure 3.1 shows the diagrams for an exemplary household of the *NoComparison* group whereas figure 3.2 shows diagrams for a *Comparison* household. Each set of diagrams contains a first (upper) diagram about the current month's consumption and a second (lower) diagram on the consumption history since September 2019. For both treatments, the savings target (reduction of 5 %) is displayed as a dashed black line. For reasons of comprehensibility, we decided to express hot water consumption in litres. *Comparison* households are additionally informed about the average consumption of households in the same room category (in grey) and the average of the best (i.e. lowest) 20% (in brown).

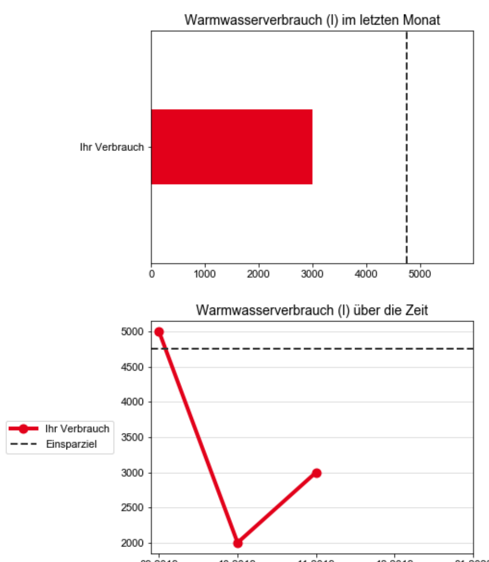

Figure 3.1: Feedback for *NoComparison*

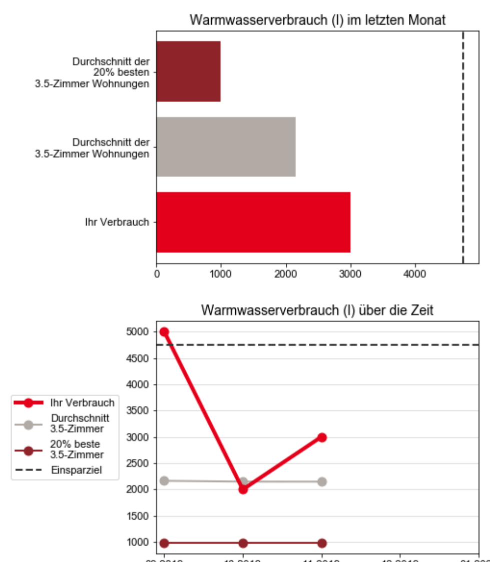

Figure 3.2: Feedback for *Comparison*

3.3 Estimation sample

The estimation sample is defined as in section 2.3.

3.4 Outcome variables

The outcome variables are defined as in section 2.4

3.5 Test of the main hypothesis

Hypothesis 1 in section 3.1 is assessed by comparing treatment groups[1] A and B (*NoComparison*) with C and D (*Comparison*). The sample is defined in subsection 3.3. The main outcome is defined in subsection 3.4. The statistical tests are equivalent to section 2.5.

3.6 Test of treatment effect heterogeneity

Hypothesis 2 in section 3.1 will be investigated as in section 2.6, with the only difference that we investigate CATEs for *Comparison* vs. *NoComparison*.

3.7 Effects over time

Effects over time will be investigated using the method described in section 2.7, but focusing on *Comparison* vs. *NoComparison*.

3.8 Exploratory analyses

In addition to the analyses specified above, exploratory analyses may be performed without pre-registration.

[1] See table 1.1 for reference.

Chapter 4

Spillovers

4.1 Research question and hypotheses

Behavioral interventions focused on one behavior may spill-over on related behaviors. Recent research implies that two different directions of spillover effects seem plausible. Cialdini et al. (1995), for instance, describe that individuals behave consistently with their initial decisions, which is classified as a *positive spillover*. Others, e.g. Merritt et al. (2010), argue that individuals may use initially 'good' behavior, to 'license' immoral behavior, which constitutes a *negative spillover*.[1]

The existence and direction of spillover effects are highly policy relevant. Positive spillovers of pro-environmental interventions provide a strong argument for the effectiveness of such instruments. Negative spillovers, however, may provoke a critical discussion on the wider suitability of behavioral instruments for sustainable policy making. To shed light on this issue, we investigate:

1. Does an intervention consisting of household-specific information about hot water consumption, hot water saving recommendations, a hot water savings target of 5%, and a lottery incentive to reach the savings target, as compared to a pure control group, lead to a reduction in hot water consumption?

2. Does this hot water specific intervention lead to a change in cold water consumption?

3. Does this hot water specific intervention lead to a change in heating energy consumption?

We hypothesize that the answer to research question 1 is 'Yes', i.e. that the intervention leads to a reduction in hot water consumption.

[1] Depending on the context, both positive and negative spillovers can result from various other mechanisms.

This section illustrates that spillovers can be positive or negative. Hence, we have no specific hypotheses regarding research questions 2 and 3. Cold water and heating energy consumption may increase, decrease, or remain unchanged.

4.2 The treatment

The treatment is the 'Hot Water Challenge' as described in section 1.2.2, with variations described in sections 2.2 and 3.2 (i.e. all treatment groups A, B, C, and D), versus the *Control* group.

4.3 Estimation sample

The estimation sample consists of all 782 buildings in the experiment, with a total of $4,775$ households. Households described in items 3, 4 (for the respective outcome), and 5 (for the respective outcome) of section 2.3 will be excluded before the building-level outcome variables are constructed.

4.4 Outcome variables

The outcome variables are defined as in section 2.4, but on the building level, indicated by b:

$$HWchange_b^{month} = \frac{HW_b^{month}}{HW_b^{baseline}} - 1 \tag{4.1}$$

The outcome variables are defined as the average percent change in hot water consumption, cold water consumption, and heating energy consumption between the baseline period (September) and the intervention periods October to January:

$$HWchange_b^{intervention} = \frac{HWchange_b^{oct} + HWchange_b^{nov} + HWchange_b^{dec} + HWchange_b^{jan}}{4} \tag{4.2}$$

$$CWchange_b^{intervention} = \frac{CWchange_b^{oct} + CWchange_b^{nov} + CWchange_b^{dec} + CWchange_b^{jan}}{4} \tag{4.3}$$

$$HEchange_b^{intervention} = \frac{HEchange_b^{oct} + HEchange_b^{nov} + HEchange_b^{dec} + HEchange_b^{jan}}{4} \tag{4.4}$$

4.5 Test of main hypotheses

Our hypotheses described in section 4.1 are assessed by comparing treatment groups[2] A, B, C, and D with the *Control* group. The sample is defined in subsection 4.3. The three outcome variables are defined in subsection 4.4. The statistical tests are equivalent to section 2.5, but on the building level.

4.6 Effects over time

Effects over time will be investigated using the method described in section 2.7, but on the building level, focusing on all treatment groups vs. *Control*, and with the three outcome variables as described in section 4.4.

4.7 Exploratory analyses

In addition to the analyses specified above, exploratory analyses may be performed without pre-registration.

[2]See table 1.1 for reference.

Bibliography

Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10):3003–37.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Brewer, N. T., Defrank, J. T., and Gilkey, M. B. (2016). Anticipated regret and health behavior: A meta-analysis. *Health Psychology*, 35(11):1264–1275.

Brook, A. (2011). Ecological footprint feedback: Motivating or discouraging? *Social Influence*, 6(2):113–128.

Cialdini, R. B., Trost, M. R., and Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, 69(2):318–328.

Costa, D. L. and Kahn, M. E. (2013). Energy Conservation Nudges and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment. *Journal of the European Economic Association*, 11(3):680–702.

Evans, K. M., Kienast, P., and Mitchell, T. R. (1988). The effects of lottery incentive programs on performance. *Journal of Organizational Behavior Management*, 9(2):113–135.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2):117–140.

Gerrard, M., Gibbons, F. X., Lane, D. J., and Stock, M. L. (2005). Smoking cessation: social comparison level predicts success for adult smokers. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, 24(6).

Hodges, J. L. and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics*, 27(2):324–335.

Landry, C., Lange, A., List, J., Price, M., and Rupp, N. (2006). Toward an understanding of the economics of charity: evidence from a field experiment. *Quarterly journal of economics*, cxxi(2):747–782.

Lange, A., List, J. A., and Price, M. K. (2007). Using lotteries to finance public goods: Theory and experimental evidence*. *International Economic Review*, 48(3):901–927.

Merritt, A. C., Effron, D. A., and Monin, B. (2010). Moral selflicensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, 4(5):344–357.

Pitman, E. (1949). *Lecture Notes on Nonparametric Statistical Inference: a course given at Columbia University, 1949*. Columbia University.

Rojo, J. and Lehmann, E. L. (2012). *Selected works of E.L. Lehmann*. Selected Works in Probability and Statistics. Springer US, New York.

Snowberg, E. and Wolfers, J. (2010). Explaining the favoritelong shot bias: Is it risk-love or misperceptions? *Current Anthropology*, 118(4):723–746.

Steiner, J. and Stewart, C. (2016). Perceiving prospects properly. *American Economic Review*, 106(7).

Suls, J., Marco, C. A., and Tobin, S. (1991). The role of temporal comparison, social comparison, and direct appraisal in the elderly's selfevaluations of health. *Journal of Applied Social Psychology*, 21(14):1125–1144.

van der Swaluw, K., Lambooij, M. S., Mathijssen, J. J., Schipper, M., Zeelenberg, M., Polder, J. J., and Prast, H. M. (2016). Design and protocol of the weight loss lottery- a cluster randomized trial. *Contemporary Clinical Trials*, 49:109 – 115.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.