Goal-Setting and Behavioral Change: Evidence from a Field Experiment on Water Conservation*

Sumit Agarwal Ximeng Fang Lorenz Goette Samuel Schoeb
Tien Foo Sing Thorsten Staake Verena Tiefenbeck Davin Wang

15 February 2022

Abstract

Modern digital technologies enable behavioral interventions through personalized feedback and goal-setting in a variety of applications. While psychologists have long argued that goals can motivate effort even when not tied to material incentives, there is little guidance on how to incorporate nonbinding goals into economic decision-making and policy intervention frameworks. We provide causal evidence on the effects of goal-setting and real-time feedback from a randomized field experiment (N = 525 households) in the context of water conservation in Singapore, using smart meters to collect fine-grained behavioral measures continuously over a duration of 4 to 6 months. Our results provide strong evidence that exogenous goals can induce significant conservation effect on top of real-time feedback if they are both challenging and attainable. The impact of goals is mostly "local": effects are particularly strong when a goal is in sight, but quickly die off when getting out of reach, suggesting a warm glow effect. Interestingly, goals seem to become less meaningful over time, although average conservation effects remain stable, which is consistent with nonbinding goals taking on the role of norms or defaults.

JEL codes: D12, D83, D91, Q41

Keywords: behavioral interventions, resource conservation, goal-setting, real-time feedback, policy interactions, randomized controlled trials

We would like to thank Café Friedrichs for providing kind hospitality and excellent cappuccino during the preparation of this manuscript.

1. Introduction

Individuals frequently act in ways that are not in line with their own values and intentions. For example, gym members want to stay in shape and healthy, yet exercise less often than they initially plan to (DellaVigna and Malmendier, 2006); students want to be well rested in the morning, yet stay up late anyway (Avery et al., 2019); entrepreneurs want to manage their businesses effectively, yet fail to follow simple rules for good financial practice (Drexler et al., 2014). One particularly relevant domain is pro-environmental behavior. Amidst growing public concern about societal challenges due to climate change and resource scarcity, many people are willing to make personal sacrifices in order to protect the environment, yet often fail to act pro-environmentally in their everyday lives (Kollmuss and Agyeman, 2002; Frederiks et al., 2015). Such intrapersonal conflicts may arise, e.g., due to lack of willpower or self-control, forgetfulness, or because the benefits of some behavior appear less immediate and salient than its costs.

Goal-setting is a simple and popular motivational tool. A large body of literature in psychology has demonstrated the motivating power of goals even when they are non-binding, i.e. there are no explicit material rewards tied to achieving or failing the goal (Locke and Latham, 1990, 2002). Similarly, the notion of "Management by Objectives" (Drucker, 1954) has been highly influential in both the theory and practice of organizational management. While economists have long studied the use (monetary) bonus incentives in organizations, they have only recently begun exploring the role of payoff-irrelevant goals, mostly in the context of self-set goals that agents can use as soft commitment tools against self-control problems (e.g. Koch and Nafziger, 2011; Harding and Hsiaw, 2014; Allen et al., 2017; Clark et al., 2020).[1]

Advances in modern digital technologies create a plethora of new opportunities for delivering simple and scalable interventions through personalized feedback and goal-setting, as they enable precise quantitative measurement of behavioral outcomes in many domains of our everyday lives, such as health behavior (Chapman et al., 2015; Edwards et al., 2016) or electricity consumption (Loock et al., 2013). Availability of large-scale fine-grained data also opens up new opportunities for behavioral scientists to evaluate the impact of different goals and to understand the underlying behavioral mechanisms.

In this paper, we provide causal evidence on the effects of goal-setting and real-time feedback from a randomized field experiment with over 2,000 individuals from 525 households in the context of everyday water conservation, using smart meters to continuously collect fine-grained behavioral measures over a duration of 4 to 6 months. We conducted the experiment in Singapore, a severely water-stressed country, where government agencies have made it a high priority to reduce daily domestic water consumption per capita

[1] Some recent studies have considered the role of non-monetary incentives to encourage effort provision in organizations, e.g. symbolic rewards (Kosfeld and Neckermann, 2011; Gallus, 2017) or tournaments without prizes (Blanes i Vidal and Nossol, 2011).

to 130 liters by 2030 (down from 141 liters in 2018), for example by promoting a wide range of water savings campaigns, often stressing that "every drop counts".[2] In our study, we target a particularly water-intensive activity, namely showering, which constitutes almost 30% of total water usage in Singaporean households (PUB, 2018a). All households were equipped with Amphiro smart shower meters that were directly installed in the shower and that automatically recorded detailed information on water usage patterns every time the shower is used. Overall, we collected data from about 320,000 shower observations over the entire course of the study.

The smart meter also allowed us to implement behavioral interventions by showing various information to subjects in real time through an integrated liquid-crystal display (Tiefenbeck et al., 2018). We randomly assigned households into one of seven experimental condition: one Control condition, one real-time feedback only condition (RTF), and five different Goal conditions. Irrespective of the condition, we programmed each device to include a baseline period of 20 showers at the beginning of the study in which it only displayed the current water temperature, which gives us a measure water consumption behavior in absence of any intervention. Thus, we have experimental treatment variation both across and within subjects.

In the Control condition, the display continued to only show the temperature information throughout the rest of the study. In constrast, from the 21st shower onwards, devices in the RTF condition started displaying in real time how many liters of water the individual is using for the current shower, thus allowing them to track their water consumption in a simple and intuitive way. In addition to real-time feedback on the absolute amount of water used, subjects in the five Goal conditions were further assigned a fixed conservation target and encouraged to keep their water usage for each shower below the respective target. The smart meters also indicated visually whether the current shower is below the target (the goal can still be achieved) or above it (the goal has been missed). However, the goal was nonbinding, i.e. there were no consequences tied to whether it was achieved or not. In a pilot study, we found that water usage per shower is roughly 20 liters on average, so we chose 10L, 15L, 20L, 25L, and 35L as possible conservation targets for the main study and randomly assigned one of these goals to each household in the Goal condition. While allowing subjects to set goals for themselves would have been an interesting extension, we focus here solely on exogenous goals in order to be able to causally estimate the effect of different goals on behavior.

Our experimental design allows us to cleanly identify the effects of real-time feedback and goal-setting on water conservation behavior by comparing outcomes across groups. In particular, it also separates the role of an exogenous goal from feedback per se. Assigning a goal is typically accompanied with feedback on one's behavior, which can already

[2] See e.g. Taylor and Accheri (2019), as well as public information provided by Singapore's National Water Agency (`pub.gov.sg/savewater`) and Government Agency (`gov.sg/features/every-drop-counts`). Accessed December 16, 2021.

have an effect of its own, as it provides information, focuses attention, and also enables individuals to set and pursue targets by themselves (e.g. Allen et al., 2017; Tiefenbeck et al., 2018). Comparing the Goal conditions with the RTF condition allows us to test in a concise way the additional impact of externally-set goals on behavior. We further generate exogenous variation in the difficulty level of the goal, ranging from very challenging (10L) to very easy (35L) for the average subject. Thus, we can evaluate the prediction from goal-setting theory that the effectiveness of goals increases in difficulty, as long as they remain realistic. Moreover, the continuous and high-frequency measurement of consumption behavior over a duration of several months gives us a sufficiently large data set to examine fine-grained behavioral responses depending on distance to the goal, as well as whether the effects of goal-setting are short-lived or remain stable over time.

Overall, the empirical results show that our interventions have a strong motivating effect on water conservation behavior. Consistent with earlier studies, we find that real-time feedback alone already leads to significant reductions in average water usage by 1.87 liters per shower relative to the Control group, which corresponds to an effect size of about 9 percent. Importantly, externally-set goals can increase conservation efforts dramatically, the reductions being twice as high (3.92 liters per shower) in the 15L condition – which turned out to be the most effective of all Goal conditions based on point estimates. However, we also find that the easist (35L) did not lead to any additional reductions in water usage compared to real-time feedback alone, with the estimated conservation effect of 1.11 liters even being somewhat smaller. In addition, the relation of goal difficulty and effort appears to be non-monotonic: while the 20L and 25L goal lead to a reduction of around 3.0 liters per shower on average, the point estimate for the most ambitious goal (10L) is 2.97 liters and thus smaller than the one for the moderately ambitious 15L goal. This non-monotonic pattern of the point estimates bears close resemblence to the conventional notion that the best goals are challenging yet attainable (Locke and Latham, 1990; Heath et al., 1999).

Furthermore, we find that goals can add motivation particularly for consumers who were already very water efficient without any intervention. While real-time feedback alone had no significant effect for consumers with below-median baseline, the 25L to 10L goal conditions induced water savings per shower of between 1.6 liters (13%) to 2.2 liters (17%) on average. In all treatment groups, the conservation effects are considerably larger for high-baseline consumers, as they have larger scope for reducing consumption, but the relative marginal benefit of externally-set conservation goals tends to be lower, as real-time feedback alone already reduces water usage by 3.25 liters per shower. Interestingly, the easy 35L condition was in fact counterproductive for this subsample of consumers, suggesting that goals may play the role of defaults or norms and potentially crowd out intrinsic motivation.

Generally, an additional implication of higher baseline usage is that a given conserva-

tion goal tends to become more challenging and less attainable. Accordingly, the pattern in heterogeneous effects for different Goal conditions matches the non-monotonic pattern in average treatment effect. For example, the interaction effects for the very easy 35L and very hard 10L goal conditions were relatively weak, which can be explained by the goal being either not challenging or not attainable, and thus irrelevant, for a significant share of individuals. Accordingly, the goal which was most effective on average (15L) also exhibited the strongst interaction effect. Non-parametric estimates of the interaction patterns suggest that objectively easier goals start perform relatively better the more water consumers consumed per shower in baseline, thus highlighting that large baseline heterogeneity may create the oppportunity to improve effectiveness by tailoring different goals to different individuals.

In addition to investigating differences in (conditional) average outcomes across experimental conditions, we make use our sample of the around 300,000 shower observations to further examine more fine-grained behavioral responses as a function of distance to the experimentally assigned goal. Based on the RTF and Control conditions, we can further construct experimental placebos to compare outcomes with and without an exogenous goal, which offers methodological advantages to studies that rely on smoothness assumptions about the counterfactual distribution (e.g. Allen et al., 2017). We observe strong bunching at goals, with the share of showers in the 0.5 liter bin below a conservation target being about 16% higher than the corresponding share in the RTF group. Using non-parametric survival analysis methods, we find that hazard rates of showers are mostly affected locally around a goal. As the amount of water used in the shower approaches the goal, the stopping rate gradually increases, peaking in the last moments before they would fail the goal. Intriguingly, we observe a sharp upward spike in the stopping probability by 44% at the very last deciliter below it. However, after the water volume has surpassed the conservation target, the hazard rate quickly drops to the level of the RTF group. This pattern of stopping probabilities strongly suggests that individuals experience a discontinuous jump in utility depending on whether they achieve the goal or not, which may be interpreted as psychological bonus reward or warm glow. In contrast, it is inconsistent with frequently used models of goals as reference points that induce loss aversion (e.g. Heath et al., 1999; Koch and Nafziger, 2011; Gómez-Miñambres, 2012), as this would predict persistently higher hazard rates once the goal is surpassed.[3]

Finally, we investigate whether the motivational power of goals is short-lived or remains stable over time. For instance, it may be the case that individuals simply become numb towards attainment or failure of nonbinding goals after some time, e.g. due to disengagement after repeated failure (Höpfner and Keith, 2021). In contrast, we find that the average conservation effects of all treatments are remarkably stable over time, with no

[3]One might argue that a model with diminishing sensitivity could also predict fading effort in the loss domain. However, even with diminishing sensitivity, local loss aversion predicts that the quitting hazard should peak after the goal is surpassed, not before.

evidence of waning (or strengthening) over a period of 4 to 6 months. Seemingly at odds with this finding, we also observe a significant decrease in bunching and goal attainment rates over the course of the study, which indicates that individuals develop a more nonchalant attitude towards the specific goal assigned at the beginning of the study. Thus, externally-set goals seem to serve as norms or default for an acceptable level of water usage per shower, with repeated experience leading individuals to form habits or adjust their expectations.

Our paper contributes to the growing literature on demand-side approaches to promote pro-environmental behavior. Behavioral interventions aimed at overcoming such barriers have been used to facilitate behavioral change in a variety of contexts such as retirement savings or public health (Thaler and Sunstein, 2008; Madrian, 2014), and are also regularly advocated as promising policy tool for fostering more environmentally sustainable household consumption behavior (e.g. Dietz et al., 2009; Allcott and Mullainathan, 2010; Reddy et al., 2017; Creutzig et al., 2018).[4] For example, influential early studies have demonstrated the impact of social-norm based interventions on household energy and water conservation (e.g. Allcott, 2011; Ayres et al., 2013; Ferraro and Price, 2013). While these interventions typically provide feedback on aggregate household consumption, recent studies have argued that interventions that enable better behavioral control and learning, e.g. through activity-specific disaggregation (Gerster et al., 2020) and higher frequency (Allcott and Rogers, 2014; Tiefenbeck et al., 2018), may increase the effectiveness. For example, in a closely related studies, Tiefenbeck et al. (2018) provide activity-specific real-time feedback in the shower through the same type of smart meter that we use in this study and document a 22% conservation effect, or, in absolute terms, a reduction of 0.6 kWh energy and 9 liters of water per shower. These results also replicate in a sample without monetary incentives and without self-selection into the study (Tiefenbeck et al., 2019). A natural question that we address is whether technology-based feedback interventions, enabled by advances in digitization and smart metering, can be enhanced by including further motivational tools like goal-setting.

Decades of studies in psychology have demonstrated the potential of nonbinding goals (or "mere" goals) for improving task performance in a large variety of contexts (Mento et al., 1987; Locke and Latham, 1990, 2002, 2019b). While economists have recently begun exploring the use of goal-setting for example to motivate healthy food choice (Samek, 2019), student performance (Dobronyi et al., 2019; Clark et al., 2020), worker effort (Corgnet et al., 2015; Brookins et al., 2017; Fan and Gómez-Miñambres, 2020), or energy conservation (Abrahamse et al., 2007; Harding and Hsiaw, 2014), there is no clear guidance yet how to incorporate nonbinding goals into economic decision-making frameworks. We

[4]Pro-environmental interventions have drawn from a broad set of instruments such as information provision, social norms, goal-setting, etc. While the general findings are that non-monetary interventions can be an effective tool in reducing energy and water usage, the quantitative effect size may be relatively small (around 2%) on average in methodologically more rigorous studies. For reviews, see e.g. Abrahamse et al. (2005), Fischer (2008), Delmas et al. (2013), Karlin et al. (2015), Andor and Fels (2018), Carlsson et al. (2021).

contribute to the literature on goal-setting by providing field evidence from a randomized experiment in a diverse sample with continuous measurement of behavior over an extended period of 4 to 6 months. While our results are consistent with many previous findings from the psychology literature — in particular that goals can motivate effort provision if they are challenging and attainable —, we further contribute to the understanding of goal-directed behavior by collecting a large data set of about 300,000 measured observations and examining fine-grained behavioral patterns in response to different goals. In line with Allen et al. (2017), who document discontinuities in the distribution of marathon finish times at round numbers (e.g. 3h, 3:30h, ...), we observe bunching of water volumes at the goal. Compared to Allen et al., our study offers methodological advances by experimentally assigning different goals to subjects and their comparing outcomes to subjects who did not receive any explicit goal. We further contribute to the literature by providing evidence on nuanced dynamic effects of repeated everyday exposure to a goal for several months.

Our empirical results also inform theoretical approaches to incorporate goals into economic models. Psychologists typically state that a goal serves as a reference standard for satisfaction (Locke and Latham, 1990).[5] This has lead Heath et al. (1999) to propose that a parsimonious way to account for many empirical regularities is that goals inherit the properties of reference points in a prospect theory value function (Kahneman and Tversky, 1979), with loss aversion and diminishing sensitivity around it. Although this view is contentious among psychologists (Locke and Latham, 2019a), it has been adopted as main modeling approach in economic studies of goal-setting (e.g. Koch and Nafziger, 2011, 2016; Gómez-Miñambres, 2012; Harding and Hsiaw, 2014; Clark et al., 2020), likely because the presence of reference dependence and loss aversion in preferences has become well-established in the economic literature by now (Della Vigna, 2009). For example, numerous studies examine whether personal earning targets influence labor supply choices (Camerer et al., 1997; Farber, 2005; Fehr and Goette, 2007; Crawford and Meng, 2011; Farber, 2015; Thakral and Tô, 2021). However, some studies have suggested that goal attainment could also be associated with a discrete jump ("notch") in the utility function Allen et al. (2017); Markle et al. (2018); Kuhn and Yu (2021) as opposed to a jump only in the *marginal* utility ("kink").[6] Our empirical results speak more in favor of a model with a discrete psychological bonus utility rather than a model of loss aversion, indicating that it may be more appropriate to interpret externally-set goals as norms or defaults rather than loss-aversion-inducing reference points.

[5]For example, (Locke and Latham, 2002) state the following: "To say that one is trying to attain a goal of X means that one will not be satisfied unless one attains X." Locke and Latham (2013) explain that "a specific, high goal eliminates ambiguity as to what constitutes high effective performance. It defines for an individual what constitutes an acceptable level of performance."

[6]Evidence on the "joy of winning" (Dohmen et al., 2011) as well as models of aspiration levels (Diecidue and van de Ven, 2006) also argue that there may be discrete rewards attached to a binary representation of success.

The remainder of the paper is structured as follows: section 2 describes the institutional details and the experimental design of the study. Section 3 provides descriptive statistics on the experimental population. Section 4 present our empirical results on average conservation effects, and Section 5 examines fine-grained responses to goals in order to better understand the underlying behavioral mechanisms. Section 6 concludes.

2. The empirical setup

In this section, we describe the randomized field experiment in Singapore, which is an island city state in South East Asia with a population of 5.54 million — and one of the most water-stressed countries in the world.

2.1. Sample recruitment and study procedures

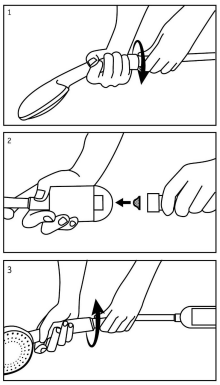
We recruited household from public housing blocks (HDBs) in 27 geographical nodes with varying population density and composition that are dispersed over the entire island and selected to create a broadly representative cross-section of Singaporean households. Appendix Figure A1 shows the geographical distribution of the participating HDB sites across the island. Note that 80% of the Singaporean population lives in HDB apartments that are built and sold by the Housing Development Board.[7]

The recruitment process was as follows: After HDBs were selected based on logistical and representativeness concerns, experimenters knocked on different flat doors — following a randomization protocol — and tried to convince households to participate in the experiment, which was framed as water conservation study. 525 households with in total over 2,000 individual household members participated in our study. All households went through informed consent procedures, and the study was approved by the IRB at the National University of Singapore. Assignment to experimental conditions was randomized within HDBs, so that we had balanced samples in each geographical node.

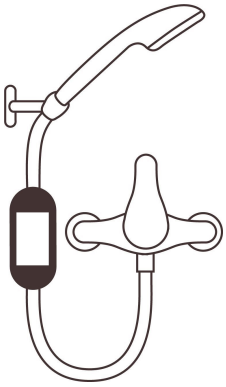
We distributed smart meters to all participating households to measure their water usage in the shower and to deliver the real-time feedback and goal-setting interventions. Deployment of the devices was carried out in June and July 2015 and the regular study duration was four months, with a subset of household (22%) being recruited for an additional 2 study months. A team of research assistants visited the households to install the devices and to explain how they work. They also interviewed one adult household member to answer a set of questions for the baseline survey. After the respective study period had ended, we revisited the households on appointment to conduct a short endline survey and to retrieve their smart meters.[8]

[7]Sources: Department of Statistics Singapore (`http://www.singstat.gov.sg`). Singapore Housing & Development Board (`http://www.hdb.gov.sg/cs/infoweb/about-us`).

[8]While direct retrievals were preferred, because we could check if devices were still installed and get a



(a) Device installation



(b) Installed device

Figure 1: Position of the Amphiro a1 smart meter in the shower.

2.2. The feedback and measurement technology

All participating households were equipped with one Amphiro a1 smart shower meter for each bathroom in their apartment. The smart meter could be easily installed between the shower head and the shower (see Figure 1), after which it measured and recorded water usage variables of every shower taken. It is small, lightweight, and powered by an integrated hydro turbine that does not noticeably affect water flow. Furthermore, it features a smartphone-sized liquid crystal display, which we programmed to show various types of information tailored specifically for the purposes of this study.

The device works as follows. Once the water flow in the shower starts, it turns on and begins to measure, among others, the water volume, water temperature, and the time passed since the beginning of the water extraction. Furthermore, its display becomes active and starts to show information. When water flow stops, the device remains powered on for three minutes to allow for short breaks e.g. for applying soap or shampoo. If water flow resumes within this three time frame, the device will continue measurement from the point where it had previously stopped. Once water flow stops for more than three minutes, the device terminates measurement, its display turns blank, and recorded information is stored as a new observation point.[9] One drawback of the lack of battery is that the device cannot keep track of global time, so that showers are only recorded in temporal order, but without time stamps.

We define "showers" as water extractions of at least 4.5 liters volume with an average flow rate of at least 2 liters per minute, whereas we classify observations as non-shower

feeling of participants' attitudes, not all of them could be reached easily and we arranged for device retrieval via postal service for 25 households.

[9]This stopping criterion introduces a small ambiguity in the measurements, as we cannot rule out that in some cases one shower is split into two, if it included a lengthy break in water flow or if two separate showers are morphed into one, e,g, when one household member uses the shower immediately after another.

(a) Mode 1 (b) Mode 2 (c) Mode 3 (d) Mode 3

Figure 2: Amphiro a1 display in different configurations

Notes: Temperature was shown only in the baseline period and for the control group. The device was installed such that the display faced directly towards the user.

water extractions otherwise.[10] The smart meters only stored detailed usage data for observations that qualify as showers according to these criteria. The reason for this restriction is that the storage capacity of each smart meter allowed for a maximum of 672 data points, which was in fact reached by 16% of the study devices. Therefore, we wanted to avoid wasting storage space for minor water extractions, e.g. for cleaning purposes. However, we did program the devices to count the total number of all water extractions as well as the cumulative amount of water used, also including extractions that did not qualify as showers. Similarly, the device stored the number of showers from the 673rd showers onward, as well as the average water volume used for these showers.

2.3. Experimental conditions

The smart meter could display tailored pieces of information to participants in real-time, i.e., while they were using their shower. For the purposes of our study, we programmed the smart meters to be in one of three possible modes, depending on the study progress and the assigned treatment. In mode 1, the device only displayed information on the current water temperature (in degrees Celsius); in mode 2, it provided real-time feedback on the absolute amount of water used for the current shower; in mode 3, it additionally provided relative feedback on water usage vis à vis a fixed conservation goal. The different display modes are illustrated in Figure 2.

Households were randomly assigned to one of seven experimental conditions and received smart meters in the respective configurations. In the control condition, subjects only ever saw information on water temperature while showering. In the real-time feedback only (RTF) condition, subjects' had devices that also displayed real-time feedback on current water usage, but did not include a conservation goal. This treatment closely

[10] A liter (L) corresponds to 33.8 fluid US ounces.

resembled the ones in Tiefenbeck et al. (2018). Furthermore, there were five different Goal conditions, in which subjects received devices that incorporated an exogenously assigned volume goal in addition to real-time feedback on absolute water usage. The goals were set at 10L, 15L, 20L, 25L, and 35L, respectively, and subjects were encouraged to keep their water consumption below this amount.[11] No explanations for the choice of the goal were provided, and the specfic goal level was only revealed with the 21st shower, when the intervention period began. From then on, it was displayed during the first ten seconds of the shower on the LCD. During showering, the display showed an injunctive message that rated the current water consumption level as “very good" if it was below 7 liters, "okay" if it was above 7 liters but below the respective conservation goal, and "too much" if it exceeded the goal.

Irrespective of the experimental condition, all devices went through a baseline period of 20 showers in which the device was in mode 1, so that it only displayed the water temperature. This allows us to collect data of baseline water consumption of households in the absence of real-time feedback or goal-setting. The interventions started with the 21st shower, from which time on participants would always see the information designated for their respective treatment group.

3. Data and descriptive statistics

Our main source of behavioral data comes from water usage measurements of over 300,000 shower observations recorded by the smart meters, representing over 2,000 individuals from more than 500 households that participated in the study. In addition, we collected supplementary survey data from households at the beginning and the end of the study, as well as from short questionnaires during the intervention. In this section, we describe our data in more detail and provide summary statistics on our experimental sample.

3.1. Water usage data

The smart meters recorded information on, among others, the water volume, water temperature, and time duration of all showers taken during the study. All but 2 of the 884 study devices we had deployed could be retrieved from the households, but for 41 devices we were not able to read out any valid data despite multiple attempts, potentially due to defective storage. Furthermore, 14 devices had no data stored at all, probably because they were never used by the households. We also have to exclude 3 households to which we had accidentally sent wrongly configured devices. Overall, we obtained valid

[11] We chose these specific targets based on data from a pilot trial with 37 households that were not part of the main study. Our aim was to be able to assign goal that ranged in difficulty level from very difficult to very easy.

Table 1: Number of observations by experimental condition

Condition	Households	Persons	Devices	Showers recorded
Control group	74	324	119 (113)	46,467 (46,405)
RTF group	70	292	110 (100)	41,967 (41,898)
10L goal group	73	312	120 (112)	44,302 (44,243)
15L goal group	72	315	117 (108)	45,601 (45,507)
20L goal group	73	313	121 (118)	49,787 (49,736)
25L goal group	74	291	118 (112)	44,787 (44,745)
35L goal group	75	303	117 (111)	47,146 (47,103)
Total	511	2,150	822 (774)	320,057 (319,637)

Notes: Underlining indicates that the number represents a lower bound, due to partially missing information for households that have not completed the baseline survey. The number of persons in a household may also include temporary or part-time residents. Numbers in brackets indicate observations for devices with at least 20 recorded showers.

water usage data for about 320,000 recorded shower observations from 822 devices and 511 households, representing over 2,000 individuals.[12] Table 1 provides an overview of the number of observations by experimental condition.

For most of our analyses, we only include devices that have recorded more than 20 shower observations, as devices with 20 observations or fewer stayed in baseline mode for the entire study and do not help us in empirically identify the effect of our interventions. Excluded devices have most likely been installed in bathrooms that are very infrequently used. Table 1 shows that this restricted analysis sample contains data from 774 devices, with the number of shower observations remaining close to 320,000. Out of these observations, 28,493 showers were recorded after the device had reached its storage limit of 672 data points. For these showers, we do not observe individual measures of water usage, but instead of this we observe the average water volume of all post-limit showers registered by a device. If not stated explicitly otherwise, we will also make use of these imputed observations for analyzing impacts on average water usage per shower.

3.2. Survey data

To supplement our behavioral data on resource use in the shower, we administered a baseline questionnaire to an adult household member when we installed the smart meters at the beginning of the study. It contained a series of items on household composition, demographic characteristics, shower habits, as well as on attitudes and beliefs towards water usage and water conservation — the latter including questions on general environmental attitude, shower comfort, and perceived water consumption ("How

[12] In 4 cases, households claimed that their device was faulty and received a replacement device. We included these households in the analysis sample, but excluded the replacement devices, because they had a second baseline period of 20 showers.

many liters of water do you think you use per shower?"). The response rate for the baseline survey was 99%.

In addition, households were asked to comple a short online follow-up survey two months after device installation, and households with study duration of six months completed an identical online survey again two months later. Finally, we conducted an in-person endline survey when retrieving the devices, whenever possible with the same individual who completed the baseline survey.[13] The follow-up and endline surveys contained questions on experiences with the shower meter, such as whether participants believed that the it was helpful, stressful, effective in changing showering behavior, and whether the goal was too difficult. In addition, they included the same set of questions about attitudes and beliefs towards water usage and water conservation as in the baseline survey. More than 95% of the households completed the follow-up and endline surveys.

3.3. Household characteristics

Descriptive statistics on household and participant characteristics are presented in Table 2 and compared to the general Singaporean population in HDB dwellings.[14] As we recruited our sample mostly from larger HDBs, the average household in our study consists of 4.2 members, while the average household size in HDBs in Singapore was 3.34 in 2015. The apartment of a modal household contained four to five bedrooms and two bathrooms. 79% of the participating households are ethnic Chinese and 12% are ethnic Indians, whereas ethnic Malay households form 5% of our sample. The composition is roughly representative of the population in Singapore, albeit with an underrepresentation of Malays relative to Chinese, Indians, and Others. The average age of individuals from participating households in our sample was 36.5 (median 35) — compared to the HDB population average of 37.9 in Singapore. About 17% of the participants were below age 15, and 10% were 65 years old or higher. Thus, our sample spans all age groups, sometimes comprising three generations within the same household, which is not uncommon in Singapore. The female share among our subjects was 53%.

3.4. Number of showers and water extractions

On average, we observe about 414 showers per bathroom over the entire 4 (to 6) months period of the trial, which corresponds to a frequency of approximately 1.3 recorded showers per person every day.[15] One concern about our intervention may be that individuals compensate shorter showers with more showers or, vice versa, that they avoid shower-

[13] 25 households sent their devices back via postal service, as we could not find a suitable retrieval appointment. In these cases, the final survey was either conducted over the phone or they filled out a paper-based survey instead.

[14] Source: Department of Statistics Singapore (singstat.gov.sg).

[15] Calculation ... The net frequency adjusted for absences may be even higher.

Table 2: Sample characteristics

Variable	Category	Frequency	Sample share	Pop. share
Apartment type	1- or 2-room	0	0%	*7.0%*
	3-room	75	14.5%	*22.8%*
	4-room	195	37.9%	*40.0%*
	5-room or EM	245	47.6%	*30.2%*
Household size	1 or 2 persons	62	12.0%	*33.8%*
	3 persons	98	19.1%	*21.5%*
	4 persons	145	28.3%	*23.2%*
	5 persons	107	20.9%	*12.6%*
	6 or more persons	101	19.7%	*8.8%*
Gender	Female	1,163	53.4%	*50.9%*
	Male	1,013	46.6%	*49.1%*
Age group	below 15 years	367	17.0%	*15.2%*
	15 - 24 years	316	14.6%	*13.0%*
	25 - 34 years	364	16.8%	*14.9%*
	35 - 44 years	338	15.6%	*15.5%*
	45 - 54 years	294	13.6%	*15.7%*
	55 - 64 years	272	12.6%	*14.0%*
	65 years and above	214	9.9%	*11.8%*
Ethnicity	Chinese	1718	78.9%	*74.3%*
	Indian	262	12.0%	*9.0%*
	Malay	101	4.6%	*13.3%*
	Other	97	4.5%	*3.3%*

Notes: Only household members for which the relevant questions in the deployment survey were answered are included. Ethnicity is assumed to be the same among all household members. Information on Singapore population obtained from the Department of Statistics (singstat.gov.sg) and from the open repository of public data (data.gov.sg) created by the Government of Singapore.

ing and thereby compromise basic hygiene needs. Furthermore, we may overestimate effects of our intervention on overall water consumption if individuals partially relocate water usage from the private shower to other facilities (e.g. wash basin, gym showers). To alleviate these concerns, we compare the total recorded number of showers per bathroom across experimental conditions in Figure 4. There is no evidence for differences in the number of showers ($p = 0.9682$). We confirm this in further robustness checks in Appendix Table A2.

Another issue could be that not all actual showers are recognized as such, because we only record detailed data for water extractions that use at least 4.5 liters. The total number of water extractions per bathroom we observe during the study period is about 510 on average. While the share of non-shower extractions seems relatively large, it should be considered that bathrooms in Singapore are often designed as closed cubicles, and that shower heads are frequently used for cleaning purposes. Still, one may be worried that our treatments had an effect along this margin, for example if individuals become more likely to take longer water flow breaks within showers in a way that a single shower

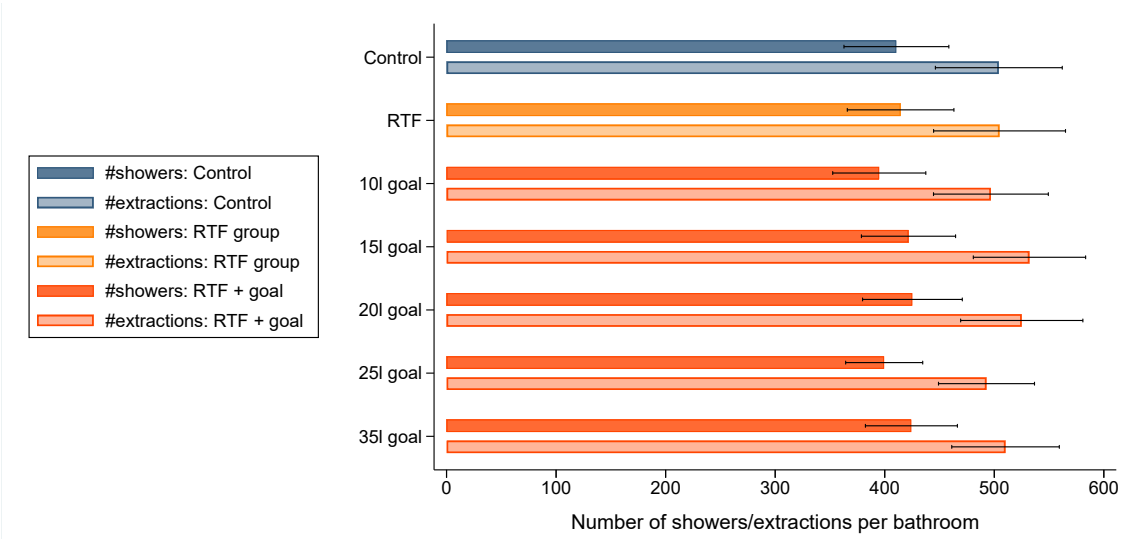


Figure 4: Number of showers and water extractions by experimental condition

Notes: Average number of shower and water extractions per device (bathroom) by experimental condition. Error whiskers represent 90% confidence intervals. Showers are defined as water extractions with at least 4.5 liters of volume.

is mistakenly recorded by the device as several extractions instead. Therefore, we additionally compare the total number of all water extractions per bathroom by treatment condition in Figure 4. Again, there are no significant differences across groups in our sample ($p = 0.9766$).

Overall, we find no evidence that our interventions induce adjustments along the extensive margin, This is important, as it allows us to make full use of the panel structure of our data and analyze (intensive-margin) water conservation effects at the level of individual shower observations rather than at the household level.

3.5. Baseline water consumption behavior

The baseline period of twenty showers per device at the beginning of the study allows us to gain insight into households' water consumption behavior in the shower in the absence of any real-time feedback or goal-setting interventions. Summary statistics are presented in Appendix Table A1. The average shower in the baseline period lasted 4.9 minutes (excluding breaks in water flow) and used up about 20.03 liters of water, which is about 50% less compared to earlier studies using the Amphiro smart meter in non-tropical countries (Tiefenbeck et al., 2018; Fang et al., 2020; Byrne et al., 2021). One reason for this is the relatively low flow rate of 4.60 liters per minute on average, which is perhaps partly due to overall lower water pressure in the high-rise HDB buildings, and partly due to the use of instant heaters as opposed to central hot water heating. Another reason is that Singapore's climate is very warm and humid, which often necessitates short showers in the middle of the day to rinse off the sweat and freshen up. This is also reflected in a low

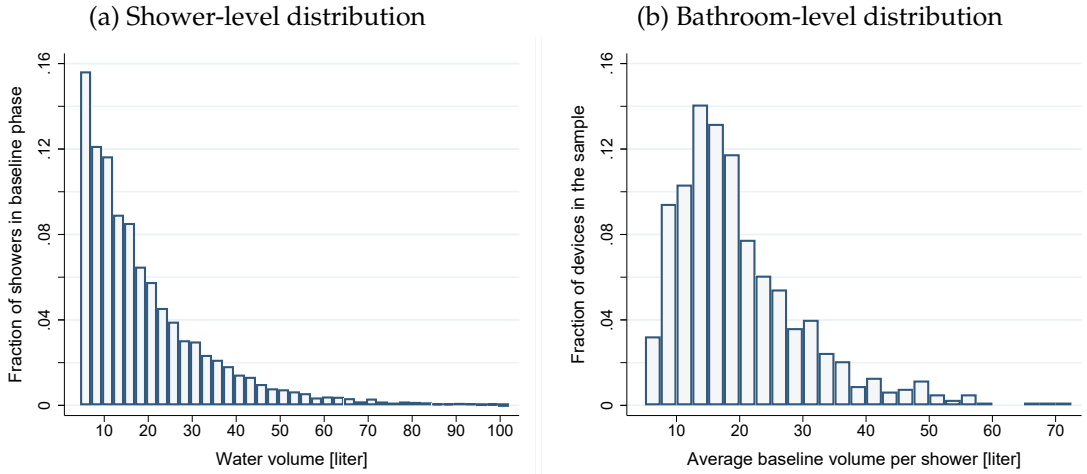


Figure 5: Baseline water usage per shower in liters

Notes: The left panel shows the full distribution of water volumes across all showers in the baseline period (first 20 showers of each device), cut off to the right at 100 liters. The right panel shows the distribution of average baseline water volume per shower at the bathroom-level.

average water temperature of 33.8 degrees Celsius and a high shower frequency of 1.3 showers per day.

Figure 5a plots the histogram of water volumes based on more than 15,000 showers in the baseline period. The distribution is heavily right-skewed, with a significant share of ultra-short showers (30.5%) that require less than 10 liters of water. The median shower only uses 14.9 liters. Hoever, there is a long tail of showers with significantly higher water consumption, with the 90th percentile lying at about 40 liters. The histogram of average shower volumes at the bathroom-level in Figure 5b shows that there is still large heterogeneity in baseline consumption behavior across houeholds and bathrooms, but the distribution becomes more concentrated and less heavily skewed, indicating substantial within-household heterogeneity of showers. Indeed, only 37.6% of the variation in baseline shower volumes is explained by across-bathroom heterogeneity. This can be driven both by differences across individuals who use the same bathroom as well by longer and shorter showers taken by the same individual. Three outlier bathrooms with an average baseline volume of more than 60 liters per shower, which can be spotted at the far end of the histogram, will be excluded for all formal analyses.

Recall that we included a diverse set of goals for the maximum water volume in our experimental design, ranging from 10L to 35L. As we can see, these goals fall into very different spots of the distribution. The 10L goal being quite ambitious for most households — only 13% of bathrooms met this goal on average even without any intervention. The 15L, 20L, and 25L goals fall into a range from moderately difficult to moderately easy, with 37% of devices registering an average baseline usage below 15 liters, and 76% below 25 liters. In contrast, the 35L goal offers virtually no challenge, as in 91% of bathrooms the

Table 3: Randomization checks

	Volume [liter]	Duration [min]	Flow rate [L/min]	Temperature [Celsius]
RTF group	0.437	0.402	-0.368	-0.336
	(1.533)	(0.300)	(0.325)	(0.353)
10L goal group	0.475	0.353	-0.269	0.245
	(1.523)	(0.291)	(0.334)	(0.312)
15L goal group	0.598	0.110	0.148	-0.549**
	(1.614)	(0.283)	(0.396)	(0.279)
20L goal group	0.147	0.163	0.152	-0.034
	(1.319)	(0.273)	(0.365)	(0.313)
25L goal group	-0.115	0.071	-0.093	-0.085
	(1.474)	(0.277)	(0.329)	(0.308)
35L goal group	1.588	0.256	0.216	-0.308
	(1.539)	(0.296)	(0.347)	(0.295)
Constant	19.400***	3.885***	5.273***	33.892***
	(1.104)	(0.208)	(0.244)	(0.209)
Observations	771	771	771	771
R^2	0.003	0.005	0.008	0.011
p-value of joint null	0.937	0.792	0.510	0.156

Only includes devices with more than 20 showers in total. Three outliers with average baseline volume of above 60 liters are dropped. Standard errors in parentheses clustered at household level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

average baseline water volume was below the goal anyway. The exogenous assignment of different goals combined with the substantial heterogeneity across households allows us to compare the impact of goal difficulty either by holding constant baseline behavior or by holding constant the goal.

3.6. Randomization checks

Our identification strategy relies on randomization producing treatment groups that are comparable with regard to observable and unobservable subject characteristics. It is naturally impossible to test the latter, but Table 3 shows good balance based on a number of key observable variables with regard to baseline behavior. Crucially, average water usage per shower is comparable across the seven experimental conditions, and a joint F-test detects no significant differences overall ($p = 0.937$). This is of particular importance as other studies generally find that households or individuals with high baseline consumption tend to respond more strongly to policy interventions (e.g. Allcott 2011; Ferraro and Miranda 2013; Tiefenbeck et al. 2018). Furthermore, there is no evidence for significant pre-intervention differences along other behavioral margins in the shower, namely duration of the shower, average water flow rate, and water temperature. While there is a single t-test that indicates signficantly lower baseline water temperature in the 15L group

relative to the Control group at the 5% level, this is in line with the rate of false positives one would expect due to multiple testing, and the F-test cannot reject the null hypothesis of joint equality across all groups ($p = 0.156$).

We further use data from the baseline survey to check for balance with regard to water conservation attitudes as well as general environmental and cost-consciousness attitudes from the baseline survey, because these could determine how individuals respond to our water conservation interventions. Appendix Table A3 shows that there are no significant differences in these attitudes across groups, further indicating that we can use the randomly assigned treatments to estimate the causal effects of real-time feedback and exogenous goals in our setting.

4. The main experimental outcomes

In this section, we present experimental results of how real-time feedback and goal interventions affect water consumption during showering on average. Furthermore, we test the stability of average treatment effects over time as well as how responses differ for subsamples of households with different baseline consumption behavior.

4.1. Descriptive evidence

In Figure 7a, we plot the moving average of water usage per shower over the course of the study. For this purpose, we construct a study progress variable that is coded to take values between 0% (beginning of the study) and 100% (end of 4-months study period).[16] Recall that in all experimental conditions, we included a baseline period of 20 showers per device at the beginning to collect behavioral data in the absence of any intervention. To clearly illustrate changes in water usage when the real-time feedback and goal-setting interventions started in the respective treatment groups, we normalize the baseline period to end at 5% study progress for all households.

The average volume per shower is about 20 liters at baseline, with a slight upward drift that continues over the entire study period in the Control group (blue line). In contrast, we observe a sharp and instant drop in water usage in the RTF and Goal conditions once the intervention started, and the conservation effects remain stable over time, with all lines following close to parallel trend. Subjects who only receive real-time feedback consistently use about 1-2 liters of water less per shower relative to the Control group. The graph also shows that the pooled Goal conditions appear to have a stronger effect than real-time information alone, as the average water volume per shower lies consistently be-

[16]Study progress of households who received the devices for six months is coded between 0 and 150. For these households, the months 5 and 6 are not presented in Figure 7a, as the trends would become very volatile due to the drastic drop in the number of observations. As the shower meter does not store global time, we construct an the measure using the order of showers and assume constant shower frequency.

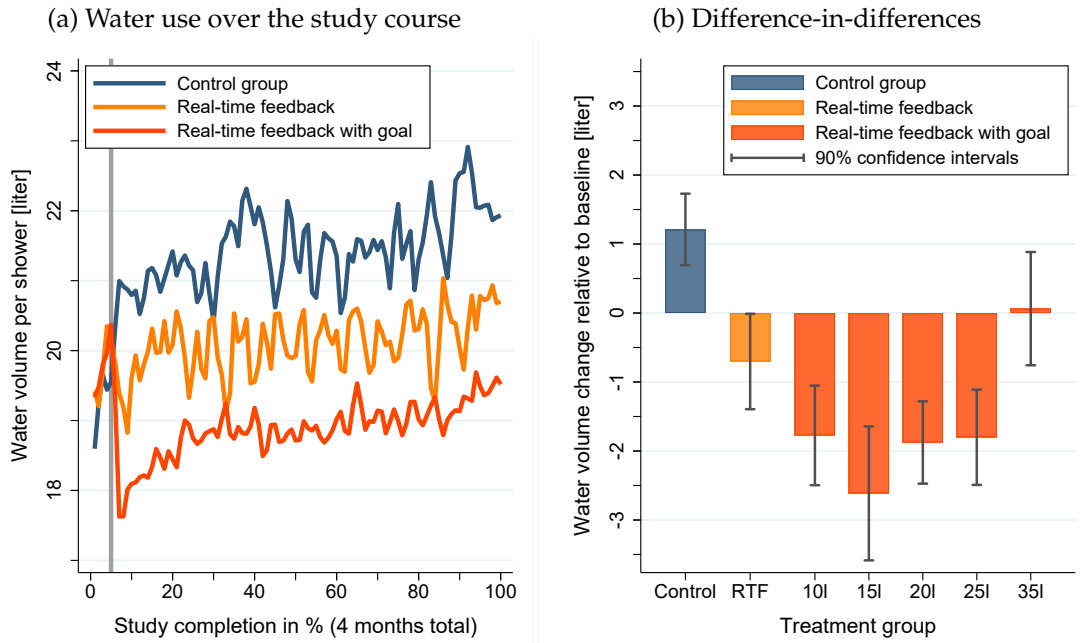


Figure 6: Descriptive evidence on the effects of our interventions

Notes: The left panel (a) shows water use over the first 4 months of the study period. Lines represent average water use at a specific study stage. Study completion percentage is defined as shower number relative to the total number of showers, where 100% spans a period 4 months. The first 20 showers of the baseline phase are normalized to 5% and the beginning of intervention is marked by the vertical black line. The right panel (b) shows changes in average water use per shower from baseline to intervention period by experimental condition. Error whiskers represent 90% confidence intervals. Both figures exclude devices with 20 or fewer recorded showers and devices with average baseline consumption of above 60 liters.

low the outcomes of the RTF condition. It is particularly noteworthy that goals and real-time feedback immediately unfold their full impact from the first shower in which they become active. This suggests that the behavioral responses are driven by higher effort or attention rather than by gradual learning about how to shower more water-efficiently.

In order to get a more accurate sense of the changes induced by the different treatments, we take the average water use for each household during the intervention phase, and subtract from it the household's average water use during the baseline phase. This reduces the number of observations to one per household, and allows us to perform a graphical difference-in-difference analysis with valid standard errors. The results are displayed in Figure 7b. The leftmost bar in the figure shows the average change in water use per shower during the intervention phase compared to the baseline phase for the control group. As was visible in Figure 7a before, there is an upward drift in the Control group of more than one liter per shower on average. By contrast, the RTF group experiences an approximately 0.7 liters decrease in water volume compared to the baseline period. The difference-in-difference estimate of the treatment effect is thus slightly below 2 liters per shower, and the 90% confidence intervals around the two means are far apart from each other, thus suggesting that the difference a strongly statistically significant.

The dark orange bars in Figure 7b represent the average changes in water volume in the five Goal conditions. They confirm the visual impression from Figure 7a that at least some goals reinforce the conservation effects compared to real-time information alone. The 15L goal shows the largest decrease in water use per shower, with a reduction that is approximately 1.5 liters higher than in the RTF condition. In addition, the pattern observed in the overall averages presents an interesting first impression of the behavioral forces at work. Remember that average water use is around 20 liters during the baseline phase. Thus, the 10L goal is relatively challenging for the average participant, whereas the 35L goal is exceedingly easy to attain. Interestingly, the moderately hard 15L goal performs somewhat better on average than the easier 20L goal or the harder 10L goal. In addition, the 35L goal clearly performs worse than any other goal condition and even worse than real-time feedback without any externally-set goal. Thus, effective goals need to be attainable but also challenging.

4.2. Average treatment effects

While the previous analyses in Figure 6 already provided descriptive evidence of the effects of real-time feedback and goals, we now exploit the full panel structure of the data to obtain more efficient estimates of the average treatment effects. We do so by estimating the following statistical model:

$$y_{is} = \alpha_i + \beta_R T_{R,is} + \beta_{10L} T_{10L,is} + \ldots + \beta_{35L} T_{35L,is} + \delta_t + \epsilon_{is} \quad (1)$$

where y_{is} is water use in shower s recorded by device i. The coefficient α_i is a device-level fixed effect that is identified through the baseline period of 20 showers at the beginning. $T_{k,is}$ are indicator variables for different treatment groups k and equal 1 if the shower occurred in the intervention phase ($s \geq 21$) and the device i belongs in the respective treatment group. The RTF group is indicated by subscript R and the Goal groups are indicated by their specific volume target (10L, 15L, 20L, 25L, 35L). The control group is omitted and serves as the reference group. Due to random assignment of households into experimental conditions, the coefficients β can be interpreted as the average treatment effects (ATE) of each treated group. We model time fixed effects by a study progress variable discussed previously, captured the coefficient δ_t for percentile t of the study duration. ϵ_{is} is the shower-specific error term. As many showers are observed for the same household on possibly up to two devices, the observations cannot be considered independent within a household. Therefore, we allow for an arbitrary covariance matrix of residuals within households by calculating heteroskedasticity-robust standard errors clustered at the household level (Abadie et al., 2017).

Table 4 column 1 presents the results for the ATEs that come from estimating the difference-in-differences model in equation 1. The coefficient estimates closely resemble

Table 4: Impact of feedback and goals on water consumption per shower

		estimating separately for three intervention periods		
	Full sample	Early	Mid	Late
	(1)	(2)	(3)	(4)
RTF group	-1.873***	-1.784***	-1.933***	-1.816***
	(0.522)	(0.495)	(0.586)	(0.615)
10l goal group	-2.972***	-2.951***	-3.126***	-2.814***
	(0.592)	(0.550)	(0.641)	(0.741)
15l goal group	-3.922***	-4.084***	-3.767***	-3.871***
	(0.661)	(0.648)	(0.714)	(0.755)
20l goal group	-3.061***	-3.185***	-2.975***	-3.032***
	(0.494)	(0.506)	(0.532)	(0.612)
25l goal group	-2.991***	-3.100***	-3.102***	-2.775***
	(0.565)	(0.537)	(0.611)	(0.674)
35l goal group	-1.108*	-1.115**	-1.088	-1.124
	(0.592)	(0.546)	(0.666)	(0.728)
Intervention	-0.260	-0.250	-0.862	0.735
	(0.381)	(0.346)	(1.172)	(1.515)
Bathroom FEs	*yes*	*yes*	*yes*	*yes*
Study progress FEs	*yes*	*yes*	*yes*	*yes*
Observations	318318	117220	117457	114461
Clusters	499	499	499	499
R^2	0.335	0.325	0.325	0.376

Standard errors in parentheses clustered at household level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

those from Figure 7b. We find that real-time feedback alone already significantly reduced water consumption by about 1.87 liters ($p < 0.001$) per shower compared to the Control condition that did not receive any feedback. This corresponds to about 9% of the baseline average, which is consistent with previous studies using the Amphiro smart meter when taking into account the baseline differences (e.g. Tiefenbeck et al., 2018; Fang et al., 2020). Crucially, we find that exogenously assigned goals can induce conservation effects above and beyond that of real-time feedback alone. For instance, the 15L goal condition reduced water usage by 3.92 liters per shower and thus by significantly more than the RTF condition ($p = 0.003$), with the estimated ATE being about twice as large. However, not all goals are created equal. While the 15L goal was the most effective goal based on the point estimates, the 10L, 20L, and 25L goal all induced a water conservation effect of around 3.0 liters and thus still performed 60% better than real-time feedback without any goal.[17] By contrast, the exceedingly easy 35L goal has does not lead to a stronger

[17]The difference in ATEs relative to the RTF group is statistically significant at the 5% level for the 15L goal group ($p = 0.024$) and at the 10% level for the 10L group ($p = 0.076$) and the 25L group ($p = 0.060$).

conservation effect than the RTF condition ($p = 0.217$), with the point estimate of -1.1 liters indicating that, if anything, it is actually less effective than having no goal assigned at all.

The empirical patterns suggests that the relationship between goal-difficulty and water conservation effort is not monotonic, but rather reverse-U shaped, which is consistent with the conventional notion that good goals should be challenging yet attainable (Locke and Latham, 1990). The easiest goal (35L) may be relatively ineffective because it offers no challenge at all for most individuals, given that the average baseline shower only used about 20 liters of water. On the other hand, most effective goal based on the point estimates is not the 10L goal, which may be unattainable for many people, but actually the 15L goal, which seems to hit a sweet spot in the trade-off between challenge and attainability. Note, though, that we cannot statistically reject the two-sided hypothesis that the 10L and the 15L goal perform equally well ($p = 0.199$), although we can strongly reject that all five goals are equally effective ($p = 0.002$).

4.3. Stability of treatment effects over time

The previous results show that, on average, suitable goals can have a strong additional effect on water conservation behavior when added to real-time feedback. Figure 7a further indicates that the effects are stable over time when pooling all five Goal conditions. However, it is conceivable that time trends vary depending on the difficulty of the goal (Goette et al., 2021). In order to examine the stability over time more formally, we split the intervention phase in three roughly equally long periods (of abut 6 weeks length) and estimate the treatment effects separately for these periods. Columns 2 to 4 in table 4 indicate a remarkable stability of effect sizes over the entire duration of the study. While the estimated coefficients exhibit some minor fluctuations over the course of several months, these differences are statistically insignificant for all treatment groups and quantitatively small, well within the range of one standard error. There is also no monotonic pattern that could indicate a clear time trend. At most, the average conservation effect of goals in our study decreases by a magnitude in the order of 0.1 to 0.3 liters per shower from the first weeks to the final weeks of the intervention.

Appendix table A4 further shows that these results are confirmed when interacting treatment effects with a four-part spline of intervention progress, so the coefficients can be interpreted as the speed with which the treatment effect changes with study progress. Two important conclusions emerge from the analyses here. First, all of our experimental treatments have an immediate effect on behavior: literally starting from the first shower of the intervention phase, the treatments are fully effective. Second, the treatment effects remain stable over our intervention period of four to six months. Therefore, there is no evidence that real-time feedback and exogenously assigned goals begin to lose their effectiveness on average conservation behavior as long as they remain in place.

4.4. Interaction with baseline usage

We continue by examining how different subgroups of individuals respond to different, randomly assigned goals. As a first step, we examine the "reduced-form" evidence on how the treatments differ in their impact as a function of the baseline water use of a household. Previous studies often find that households or individuals with high baseline consumption tend to respond more strongly to policy interventions targeted at their conservation behavior (e.g. Allcott 2011; Ferraro and Miranda 2013; Tiefenbeck et al. 2018). For example, Allcott (2011) reports that Opower home energy reports achieved virtually no savings for households in the bottom decile of baseline energy use, whereas the treatment effect for top-decile users was 6.3% savings. Tiefenbeck et al. (2018) estimate that real-time feedback has an additional conservation effect of 0.31 kWh for a 1 kWh increase in baseline energy use per shower. One straightforward way to interpret this is that high-baseline users have higher scope for reducing their consumption. The assignment of goals adds an additional dimension, as holding constant the specific conservation target, e.g. 15 liters, higher baseline consumption level implies a higher difficulty of the goal. Non-monotonicities in the response to goal difficulty would therefore also be reflected in differential responses of high- and low-baseline users to our intervention.

We analyze heterogeneity by baseline consumption first by splitting the sample into consumers with average baseline water use per shower above and below the sample median (17.4 liters), respectively. Second, we also estimate an interacted model

$$\begin{aligned} y_{is} = \alpha_i + \beta_{10L}\, T_{10L,is} + \ldots + \beta_{35L}\, T_{35L,is} + \beta_R T_{R,is} + \gamma_C I(s \geq 21) \times z_{it} \\ + \gamma_{10L} T_{10L,is} \times z_{it} + \ldots + \gamma_{35L}\, T_{35L,is} \times z_{it} + \gamma_{RT} T_{35L,is} \times z_{it} + \delta_t + \epsilon_{is} \end{aligned} \tag{2}$$

where the treatment indicators are interacted with baseline consumption z_i, i.e. average water use during the baseline phase for each household. Notice that even though we have fixed effects in place, we need to allow for a main effect interacting the intervention indicator with z_i, because there could be differential trends associated with different values of z_i, for example due to mean reversion or other baseline-dependent serial correlation . These will be captured by the coefficient γ_C.

The results are displayed in Table 5. Columns (1) and (2) show the estimated treatment effects for below-median and above-median consumers, respectively, and column (3) shows the estimated interaction effects in the linear interactions model from equation 2. Consistent with previous literature, we observe that conservation effects are significantly stronger for subjects with high baseline consumption. Real-time feedback alone had no significant effect for low-baseline consumers, who used only 12.49 liters per shower on average in the baseline phase, whereas it reduced water use per shower by 3.25 liters on average for high-baseline consumers, who used 27.18 liters per shower on average in the baseline phase. This is also reflected in an estimated linear interaction of

Table 5: Heterogeneous effects by baseline water consumption

	Median split		
	low users (1)	high users (2)	linear interactions (3)
RTF group	-0.383	-3.251***	-0.235***
	(0.628)	(0.843)	(0.056)
10l goal group	-2.166***	-3.620***	-0.122**
	(0.624)	(0.940)	(0.059)
15l goal group	-1.855***	-6.028***	-0.354***
	(0.548)	(1.105)	(0.078)
20l goal group	-1.585***	-4.157***	-0.260***
	(0.545)	(0.771)	(0.068)
25l goal group	-1.598***	-4.621***	-0.251***
	(0.559)	(0.985)	(0.069)
35l goal group	-0.635	-1.426	-0.049
	(0.581)	(0.948)	(0.089)
Baseline	–	–	0.010
			(0.039)
Main treatment indicators	*n/a*	*n/a*	*yes*
Bathroom fixed effects	*yes*	*yes*	*yes*
Study completion fixed effects	*yes*	*yes*	*yes*
Observations	147837	170481	318318
Clusters	305	310	498
R^2	0.170	0.242	0.336

Columns (1) and (2) estimate equation 1 for the subsamples of devices with below- and above-median baseline consumption, respectively. Column (3) shows the coefficients for interaction effects from estimating equation 2. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

0.235 liters lower consumption per 1 liter increase in baseline consumption in the RTF group. Note that the relative average treatment effect of real-time feedback was around 9%, hence higher baseline consumption is associated with an overproportional increase in effectiveness, as found in several previous studies of resource conservation (see, e.g., Allcott, 2011; Allcott and Rogers, 2014; Tiefenbeck et al., 2018).

The treatment effects for the goal conditions exhibit qualitatively similar interactions with baseline consumption, but there is also significant variation in the extent of heterogeneity induced by different goal difficulty level. Indeed, we can rule out at the 1% level that the interaction effects in column (3) are equal among all five goal conditions ($p = 0.0084$). Column (1) shows that even in the subsample of low-baseline consumers, where real-time feedback alone was ineffective, all goal conditions except for the 35L group induced statistically significant conservation effect of 1.59 to 2.17 liters per shower,

which is equivalent to 13% to 17% of baseline consumption. Although we cannot reject the null hypethesis of equal effects in the four goal conditions ($p = 0.651$), it is worth noting that the point estimate is largest for the most difficult 10L goal, which achieved a reduction in water consumption by 2.17 liters, which is significantly more than in the RTF condition ($p = 0.0032$). However, despite its impressive performance among low-baseline users, the effect of the 10L condition in the subsample of high-baseline users (-3.62 liters) was comparable to that of the RTF condition ($p = 0.7140$) — accordingly, its linear interaction coefficient in column (3) is also closer to zero ($\beta_{10L} - \beta_R = .1131$, $p = 0.058$). This finding is consistent with the theoretical prediction that a goal so difficult that it becomes unattainable does not have strong effects. As baseline consumption of a household increases, attaining the 10 liter goal becomes subjectively harder and harder, thus its additional motivational power eventually vanishes. At the other extreme, the most easy 35 liter goal had no significant conservation effect for low-baseline users, as expected, because it is not challenging and thus likely simply ignored.[18] Perhaps more surprisingly, the 35L condition was in fact less effective than the RTF condition for high-baseline users ($\beta_{35} - \beta_R = 1.826$, $p = 0.0723$), which may be suggestive evidence for boomerang effects or crowding out of intrinsic motivation to reduce water consumption in response to feedback. As a consequence, the interaction with baseline consumption is very low and insignificant.

The effect heterogeneity across baseline use is strongest for the intermediate 15L, 20L, and 25L goals, which is consistent with behavioral predictions based on goal-setting theory and the warm-glow model, in which effective goals need to be both challenging and attainable. Intuitively, in a heterogenous population, an increase in baseline consumption at the top level first induces stronger behavioral responses, because the goal becomes subjectively more challenging; at the same time, it still remains attainable once moving into the bottom level. In contrast, a goal that is on average very difficult becomes unattainable for individuals at the bottom, whereas a goal that is too easy becomes unchallenging for individuals at the top. In line with this reasoning, the interaction effect is quantitatively largest for the 15L condition, which also had the quantitatively strongest ATE, as it seems to embody a sweet spot in the trade-off between challenge and attainability. We estimate that for every one liter increase in the baseline consumption, the treatment effect increases by 0.354 liters in this condition, whereas the coefficients for the 20L and 25L groups are 0.260 and 0.251 and thus very similar as for the RTF group.

In Figure 8, we further illustrate the relationship between behavioral responses and baseline consumption in a nonparametric way by estimating local linear regressions at the bathroom level for each experimental condition separately. Note that we cut off the graph to the right, because the confidence bands for devices with the highest baseline

[18] For subjects with below-median water consumption per shower, only 1.83% of showers in the baseline phase used up 35 liters of water or more. Even for above-median users, a 35 liter shower lies approximately in the 75th percentile of the baseline distribution.

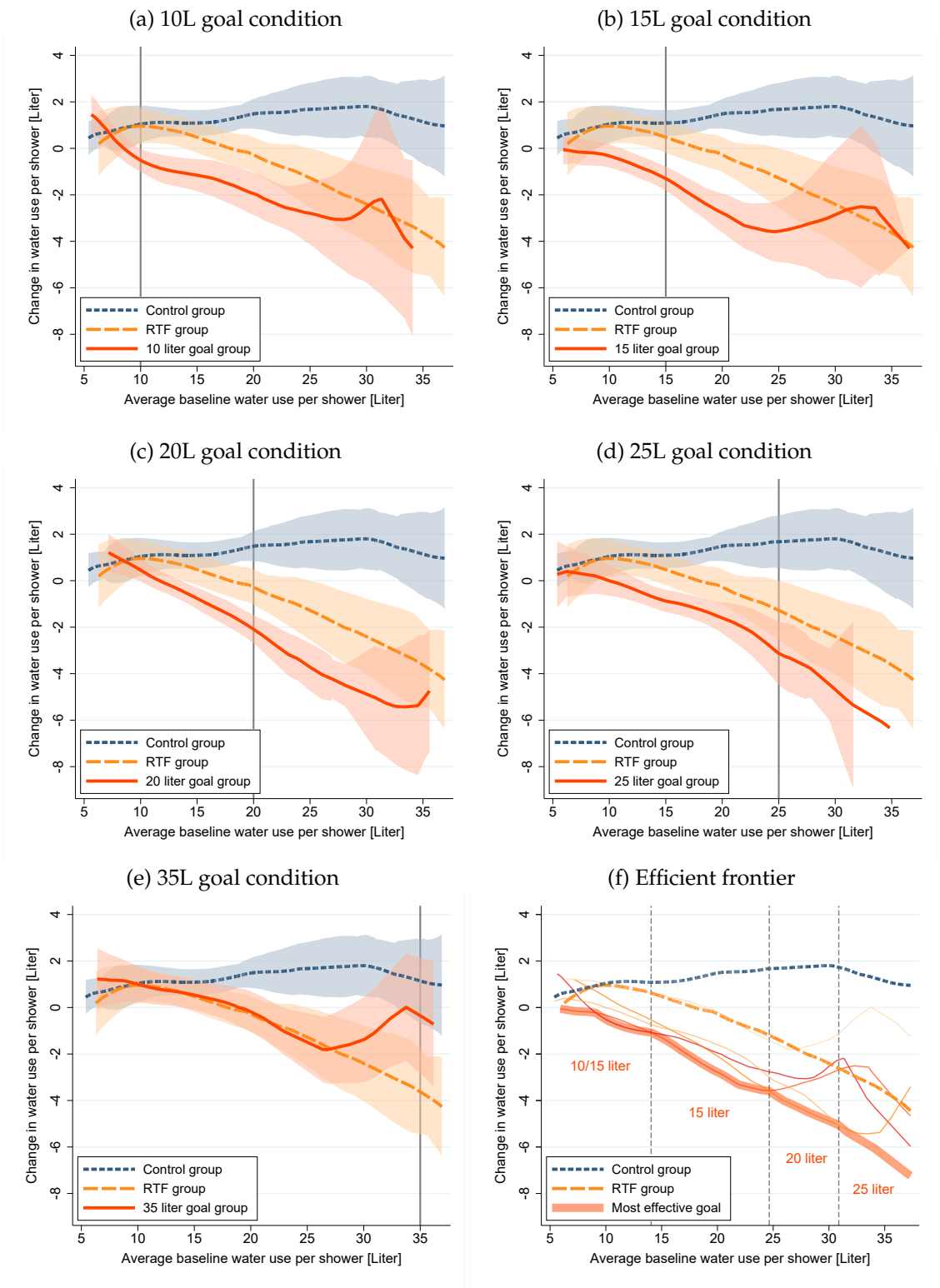


Figure 8: Local linear regressions of DiD estimates by baseline consumption

Notes: All figures present results from local linear regressions on bathroom-level using the Epanechnikov kernel with a bandwidth of 4. The outcome variable is the change in average water consumption per shower from baseline to intervention period. The independent variable is the average water consumption per shower in the baseline period. Shaded areas represent 90% confidence bands. Devices with average baseline consumption of more than 37 liters (93.4th percentile) are not displayed for visual reasons, as the confidence intervals become very wide due to small local sample sizes and large noise.

usage become very wide. The local linear estimates confirm the results in Table 5 that real-time feedback is mostly ineffective for consumers who were already very water-efficient, but starts to become effective for households with an average baseline usage of above 15-20 liters, with the water conservation effect now increasing approximately linearly compared to the control group. For the Goal conditions, the pattern is in principle similar, but varies across difficulty levels. In the 10L and 15L goal conditions, even households with low baseline usage of around 10 liters per shower already show relatively large conservation effects, but the effect estimates converge to those of the RTF condition for high-baseline consumers, as the goals become too challenging, Indeed, the slope is almost generally flatter in the 10L condition compared to the RTF condition. In contrast, the estimates for the 15L condition exhibit a steeper slope in the range between 10 liters and 25 liters baseline usage, which is where the majority of households fall into (see Figure 5b). The estimates for the 20L and 25L goal conditions roughly resemble the estimates for the RTF group with a downward parallel shift, whereas the local effects in the 35L goal group are almost identical to the RTF group except for high baseline users, for which real-time feedback without goals is actually more effective.

Figure 8f compiles the nonparametric fits for all treatment groups in a single graph, which allows us to trace out the treatment effect "frontier" based on the most effective goal (based on the point estimates in our sample) as a function of baseline consumption. A highly suggestive pattern arises: at the lower end of the baseline distribution, the 10L and 15L goal conditions induce the largest conservation effects; in the middle of the distribution, where the largest share of households fall into, the 15L goal performs best; at the higher end of the distribution, the 20L condition and 25L condition start surpassing it. This pattern again supports the notion that moderately challenging goals are most motivation-enhancing, where the optimal goal may vary across individuals due to differences in subjective difficulty levels. However, the most easy 35 liter goal breaks with the pattern to a certain degree, as it seems to become counterproductive exactly for the subset of households for whom achieving it is not a sure-fire endeavor anymore. This could be explained in a way that externally-set goals also represent a type of socially acceptable standard, which may crowd out potentially more ambitious personal standards.

5. Behavioral mechanisms of goal-setting

The results in the previous section show that goals add a stable motivation to water conservation efforts on top of real-time feedback. This section discusses the evidence on the predictions by the loss-aversion (LA) and fixed-penalty (FP) models, and assesses which of the two models can better account for the evidence.

5.1. Excess mass at the goals

The previous analysis examined how conditional means in water conservation outcomes changed as a function of the experimental conditions and across various subgroups. As a final step to better understand the behavioral mechanism underlying the motivating effects of goal-setting, we leverage the large sample size of around 300,000 total recorded shower observations to conduct more fine-grained analyses of treatment responses at the individual shower level.

We do so by first exploiting the random assignment into experimental conditions to compare the empirical distributions of showers in the intervention phase between the goal groups and the RTF group. If conservation goals serve as reference points for evaluating success and failure, e.g. by creating a kink (loss aversion) or a notch (fixed reward) in the utility function, then we would expect a general shift in probability mass from above the goal to below the goal, and specifically also bunching of outcomes at the respective goal (Kleven, 2016). For example, Allen et al. (2017) provide evidence that the distribution of marathon runners' finish times exhibits excess mass below and missing mass above round numbers (e.g. 3 hours, 4 hours).

The advantage of our setting is that we have experimentally-induced variation in both whether households receive a goal at all and what the specific goal is, and thus do not need to rely on smoothness and local boundedness assumptions to construct a counterfactual distribution. Still, we need to account for the fact that the goal group receives feedback on water use, and, e.g., individuals may have a higher likelihood of ending a shower at, e.g., 20L even in the absence of any goal; second, we are using goal distance as independent variable. Since goals differ across the five goal conditions, the question arises of how to construct a counterfactual with the same conditional water consumption but not subject to a goal. In order to address the first issue, we choose the RTF condition as our counterfactual group, thus holding all effects from feedback on the distribution constant. In order to construct a group with comparable conditional water use, we construct the the counterfactual distribution as a function of a "placebo" goal distance, in which we use each observation from the real-time condition five times, to calculate the share of showers for each of the placebo goals from the goal conditions.

In Figure 9, we group shower observations during the intervention period into 1 liter bins based on their distance to the respective goal and plot the excess and missing mass of showers in goal group versus RTF group households. The visual impression is striking. Assignment of an exogenous conservation goal induces a consistent shift in probability mass from above to below the goal, thus providing compelling evidence that individuals exert effort in order to avoid exceeding the target level that was externally assigned to them. Moreover, the shifts in the empirical density function are not uniform. There is strong bunching in the 1 liter bin just before the respective goal, with showers in the goal conditions have a 0.68%p higher probability to fall into this bin, which corresponds to a

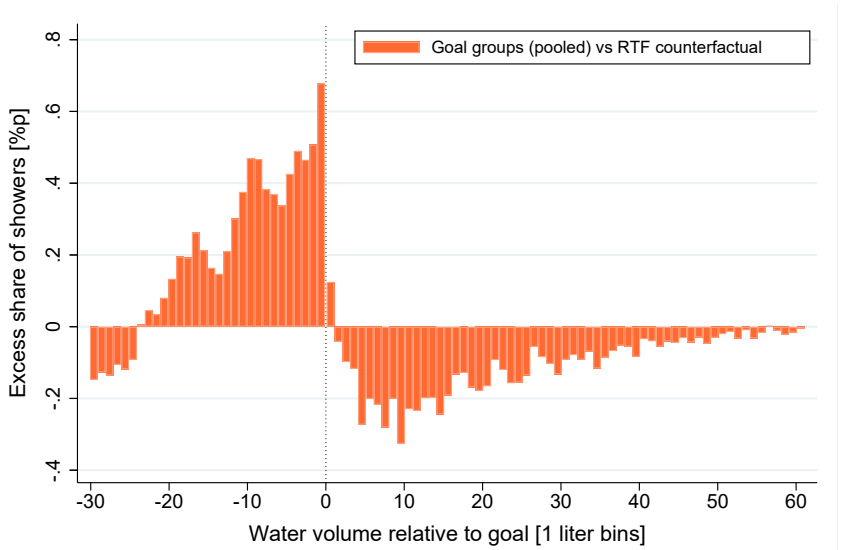


Figure 9: Excess share of showers in the goal groups relative to the RTF group

Notes: Bars are the difference between the share of intervention phase showers falling into a respective water volume bin in the goal conditions versus the RTF condition

relative increase in 25% compared the respective share of showers in the RTF condition (2.7%).[19] The spike in distribution just before the goal is followed by a sharp drop in the relative share of showers just above the goal, although still remaining slightly higher than in the absence of an explicit goal. The largest amount of missing mass is found at about 5 to 10 liters above the goal, after which the distributions converge again at a slow rate. While bunching is most evident just below the goal, there is an excess mass of showers up to 20 liters below the goal relative to the RTF condition, which suggests that the influence of goal-setting on consumption behavior is not limited to extremely local responses around the goal. Note that due to the water volume of a shower being bounded from below by 4.5 liters, each goal condition is only represented from $-G + 4.5$ onwards in Figure 9, where G is the conservation target. Appendix Figure A4 compares the distribution of each goal condition separately with the distribution in the RTF group. Interestingly, we observe missing mass of ultra-short showers in the 35L group, which again suggests a boomerang effect for very easy goals.

To estimate local bunching around conservation goals more formally, we estimate a linear probabilities model with an indicator for a shower falling in a particular volume bin Δ_V close to a salient thresholds V (e.g. 10 liters, 15 liters, ...) as dependent variable:

$$\mathbb{1}\{y_{is} \in \Delta_V\} = \alpha_i + \beta_1 T_{is} + \left(\beta_2 + \gamma \cdot \mathbb{1}_i\{V = G\}\right) T_{is}^{goal} + \delta_s + \theta_V + \epsilon_{is}\,. \qquad (3)$$

We include fixed effects for bathroom (α_i), intervention period (δ_s), and threshold (θ_V). T_{is} is an indicator for a shower by a treated household (both RTF and goal groups) in the intervention period, and T_{is}^{goal} is an indicator for intervention period showers by house-

[19] The distribution of showers with regard to goal distance are presented in Appendix 6 Figure A2.

Table 6: Probability of showers just above or below a salient threshold

	below salient threshold			above salient threshold		
	0.5L bin	1L bin	2L bin	0.5L bin	1L bin	2L bin
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.008	0.002	0.014	0.005	0.007	-0.022
	(0.010)	(0.012)	(0.015)	(0.010)	(0.011)	(0.014)
Treated × goal group	-0.010	-0.006	-0.014	-0.006	-0.011	0.008
	(0.008)	(0.010)	(0.012)	(0.008)	(0.008)	(0.011)
Matching goal	0.026***	0.031***	0.034***	0.000	-0.006*	-0.025***
	(0.005)	(0.005)	(0.007)	(0.003)	(0.004)	(0.007)
Intervention period	0.013**	-0.000	0.001	0.001	0.002	0.012
	(0.006)	(0.008)	(0.010)	(0.006)	(0.008)	(0.010)
Constant	0.095***	0.159***	0.322***	0.088***	0.145***	0.282***
	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)
Bathroom fixed effects	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Threshold fixed effects	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
N	289710	289710	289710	289710	289710	289710
R^2	0.039	0.064	0.158	0.033	0.057	0.128

Notes. Results come from estimating equation 3 using ordinary least squars. The dependent variable is an indicator for whether a shower falls into a particular volume bin around a salient threshold. We consider thresholds in steps of 5 from 10 liters to 45 liters. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

holds in one of the five goal conditions. We additionally interact the latter with a match indicator that takes the value 1 only if the threshold coincides with the conservation goal. The coefficients β_1 and β_2 indicate whether households who receive real-time feedback use salient numbers as anchor that do not correspond to an externally-assigned conservation target. The main coefficient of interest γ captures how much more likely it is that showers fall in a certain bin close to the goal relative to other salient thresholds. Thus, we exploit the random assignment of different goals to households to identify changes in the local distribution around a goal.

We consider three bin sizes $|\Delta_V| \in \{0.5L, 1L, 2L\}$ and estimate equation 3 separately for these bins above and below the thresholds. Table 6 presents the results of this empirical exercise. Columns (1) to (3) show that there is signficant bunching of showers at the goal. For example, showers were 3.1 percentage points (19%) more likely to be placed less than 1 liter below a threshold that corresponds to an exogenous goal, and 2.6 percentage points (27%) more likely to be placed less than half a liter below a goal. On the other hand, there are only quantitatively weak signs of missing mass up to 1 liter above a goal threshold, but the share of showers that are up to 2 liters above a goal is 2.5 percentage points (9%) lower. In contrast to Allen et al. (2017), we find no evidence of strong local responses to salient thresholds that are not associated with an explicit external goal. However, this does not necessarily imply that individuals who receive real-time feedback do not attempt to set and achieve personal conservation goals, as discontinuities in the

distribution may as well be hidden by heterogeneity in self-set goals.

While models of goals as notches or kinks in the utility function both predict bunching at the goal and missing mass above it (Kleven, 2016; Allen et al., 2017), the specific patterns — in particular the gradual build-up in excess mass starting far below the goal, as well as the gradual manifestation of missing mass above the goal — are at odds with a simple model without optimization frictions, but can potentially be explained by the presence of inattention or uncertainty (Kleven and Waseem, 2013).

In general, however, inferring local behavioral responses from excess mass in the empirical probability density functions can be partly complicated due to broader shifts in the cumulative density of water consumption levels in response to feedback and goals. It can thus be hard to interpret excess mass in a certain range, as it could be driven both by a local change in the probability of stopping a shower or a general shift of high-volume showers to showers with lower volume. Therefore, in the next step, we examine the stopping probabilities of individual shower in terms of the hazard rate, i.e. the probability that a shower stops at a given water consumption level *conditional* on "surviving" until this point.

5.2. Goal distance and stopping hazards

To give a graphical overview of how goals affect the stopping probabilities of showers in Figure 10, we again pool all five goal conditions and calculate the hazard rate as a function of the distance to the goal in steps of deciliters, our most fine-grained unit of measurement. The hazard rate at point k is defined as the conditional probability of stopping between $k-1$ and k deciliters relative to the goal, given that the relative water volume is above $k-1$ deciliters. Hence, a higher hazard rate reflects a higher probability to end the shower at a given point and thus higher effort to conserve water, irrespective of where k lies in the distribution. As before, we construct the counterfactual hazard rate for the goal groups by assigning placebo goals to each observation in the RTF group five times. To flexibly control for baseline differences across experimental conditions, we further adjust the hazard rates in the intervention period by dividing through local linear estimates of the baseline hazard ratio between the goal groups and the RTF group.[20] Thus, the following results can be interpreted as difference-in-differences of hazard rates.

Figure 10a plots the hazard rates as a function of water volume relative to the conservation goal in deciliters, as well as smoothed estimates using local linear regressions. In addition, Figure 10b plots the hazard ratio relative to the RTF counterfactual using the smoothed hazard rate estimates, with pointwise confidence intervals obtained from a

[20]More specifically, we run separate local linear regressions of the baseline hazard rates by goal distance for the goal conditions and the RTF condition with Placebo goals. We then use the smoothed estimates to calculate the local hazard ratios and divide the intervention period hazard rates in the goal conditions by the respective hazard ratio. The results without any adjustment for baseline differences can be found in Appendix Figure A5 and look very similar.

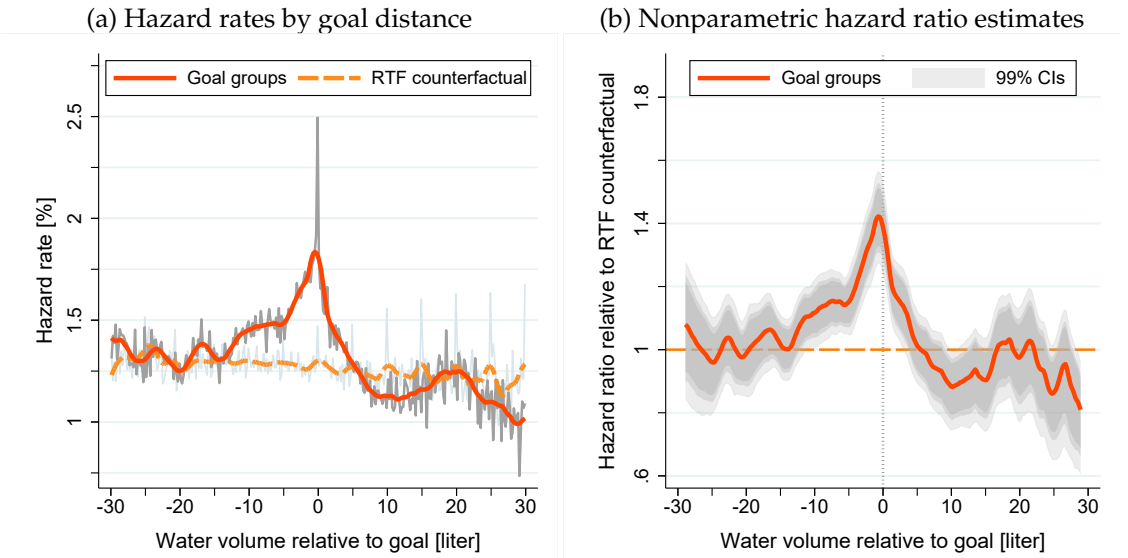


Figure 10: Stopping hazard by goal distance

Notes: The left panel (a) plots the hazard rates of showers by deciliter of distance to the respective conservation goal. Smoothed estimates are obtained through local linear regressions using the Epanechnikov kernel with bandwidth 0.6 liters. Hazard rates are adjusted for baseline differences between experimental conditions by dividing through smoothed local estimates of the hazard ratio between goal and RTF condition in the baseline period (see footnote 20). The right panel (b) plots the hazard ratio, calculated from the ratio of smoothed hazard rate estimates in the goal group and the RTF counterfactual. Bootstrap percentile confidence intervals are obtained from clustered bootstrapping with 4,000 simulations, using households as unit of resampling. Different shades of grey reflect 90%, 95%, and 99% confidence regions, respectively.

block bootstrapping procedure that accounts for clustering at household level.[21] The counterfactual hazard rate stays relatively constant, fluctuating around 1.25% with a slight downward trend. Some wave-like patterns with humps at round numbers hint at the presence of self-set goals à la Allen et al. (2017), but are too small to be detected in Table 6. In comparison, the hazard rates in the goal conditions show a very clear pattern. Stopping behavior is relatively unaffected by the exogenously assigned goal when the water volume is still more than 15 liters below the goal, as there is large remaining scope for finishing the shower in time.[22] As individuals approach the goal, the hazard rate increases above the counterfactual rate and reaches its peak just below the goal. What springs to the eye is the enormous spike in the stopping probability at the very last deciliter, which jumps from about 1.75% up to 2.5% and then immediately down again. While the smoothed estimates generally track the movements of the empirical hazard rate of the goal conditions very well — capturing about 84% of the variation within 30 liters around the goal — they fail to account for the anomalous spike at the goal.[23] This is per-

[21] Specifically, we resample households 4,000 times and ... we intentionally undersmooth the local linear hazard rate estimates ... obtain equal-tailed percentile confidence intervals.

[22] There are also some noticeable ups and downs in the goal condition hazard rate below the goal. These wave-like patterns may driven by the subgoal at 7 liters at which point the injunctive message switches from "very good" to "okay", as the humps tend to coincide with $7 - G$.

[23] We can quantify the anomaly by fitting a local linear estimate that uses all empirical hazard rates except for the one at the last deciliter before the goal, in the spirit of the bunching estimator approach by (Chetty et

haps the single most powerful piece of evidence in this study that individuals respond to nonbinding, exogenously-assigned goals. Interestingly, the hazard ratio rapidly reverts and becomes statistically indistinguishable from 1 after just three to four liters since the goal has been missed, even dropping below 1 for showers with higher water volumes, which stands in contradiction to loss aversion models, which would predict *higher* stopping rates in the loss domain, i.e. when the goal has been missed, compared to the gain domain.

This setup allows us to test the predictions of the loss-aversion and fixed-penalty model from a different angle. If loss aversion is driving goal effects, then quitting hazards should be unaffected (up to some uncertainty owing to randomness in stopping) before water usage has reached the goal. The stopping hazard should increase once the individual is past the goal and in the loss domain with the correspondingly higher marginal disutility from water use. By contrast, the fixed-penalty model implies that stopping hazards should be higher as the individual approaches the goal. Since the penalty is fixed and incurred as the individual surpasses the goal, the individual has an incentive to stop somewhat early owing to the randomness in the water used.[24]

This pattern is fully consistent with the fixed-penalty model: individuals stop somewhat ahead of the goal in order to avoid overshooting due to randomness. However, once they overshot, goal-related efforts to stop vanish and the stopping hazard becomes indistinguishable from the one of the real-time feedback group that was not assigned a goal. At the same time, the pattern is difficult to reconcile with the loss-aversion model, in which the higher marginal disutility from surpassing the goal motivates stopping efforts, as the stopping hazard in the goal conditions quickly reverts to the one of the real-time group once the goal is missed.

5.3. Changes in behavioral response over time?

The underlying behavioral mechanism of how goals enter the utility function also has implications of the stability of the treatment effects over time. If one takes the view that goals take on the role of reference points directly (See, e.g., Heath et al., 1999), then responses should remain stable over time. However, in a model of expectation-based reference points Koszegi and Rabin (2006, 2009), it is possible that goals may not only affect reference points directly, but also shift expectations.[25] In such a model, a shift in expectations can be self-fulfilling and subsequently affect behavior. However, this raises the

al., 2011). Comparing the actual hazard rate to the leave-one-out estimate indicates a discontinuous jump by 0.76 percentage points at the goal, which corresponds to about 44%. Using clustered Monte Carlo bootstrap inference, we can show that this jump is highly statistically significant, as in 4,000 bootstrap simulations there was not a single instance in which no large positive spike in the hazard rate occurred.

[24] If there were no randomness in water use, the model would predict bunching at exactly the goal.

[25] The evidence from lab experiments with regard to the expectations mechanism is mixed. While some papers find evidence of the comparative statics predictions (Abeler et al., 2011; Ericson and Fuster, 2011; Goette et al., 2020), and others rejecting its predictions (Gneezy et al., 2017; Cerulli-Harms et al., 2019)

question of whether the impact of goals becomes less effective over time. Suppose an individual was assigned a hard goal (compared to her baseline water use). If this affects her expectations and thus her reference point, both of the models outlined above would predict an increased conservation effort. However, as time goes by and the individual repeatedly falls short of the goal, this may affect her expectations, and thus her reference point. Thus, it is possible that goal effects are temporary and gradually losing their effect on behavior.

In Figure 11, we further split the data into the three phases of the intervention period to examine whether behavioral responses adjust over the course of several months. The first observation is that stopping hazards for individuals who only received real-time feedback remain fairly stable, mirroring the results for average water consumption from Table 4. The second observation is that, qualitatively, the pattern induced by exogenous goals also remains similar in the later phases of the intervention, with stopping hazards gradually increasing starting from 10 to 15 liters below the goal, peaking with an anomalous spike at the goal, and then quickly plummeting again. However, the third observation is that, quantitatively, the peak at the goal diminishes considerably in magnitude over time. In the first weeks of the intervention, the hazard rates exhibits an impressive jump by 53% (0.96 percentage points) to 2.76% at the goal, whereas in the final weeks it "only"goes up by 38% (0.63 percentage points) to 2.26%. We corroborate this finding in Appendix Table A5, which extends the analysis in Table 6 by an interaction with study progress and shows that bunching of showers in the 0.5 liter and 1 liter bins (but not the 2 liter bin) below a goal decreases significantly over time, with point estimates implying that the excess mass vanishes competely after approximately 6 months.

While we have shown previously in section 4.3 that the average water conservation effects induced by nonbinding goals remain largely stable over the duration of our study, our data paints a more nuanced picture when also considering the set of results in this section. There are two possible explanations. First, it may be the case that individuals become comfortably numb towards the externally-set goal over time, as they develop a more nonchalant attitude towards achieving or missing it; still, they continue to use lower amounts of water due to, e.g., habit formation (Charness and Gneezy, 2009; Wood and Rünger, 2016; Byrne et al., 2021) or endogenously adjusting reference points (Koszegi and Rabin, 2006, 2009; Thakral and Tô, 2021). Second, it is possible that individuals continue striving to achieve the specific goal that was assigned to them at the beginning of the study, but learn to become more proficient at predicting and regulating their water usage and thereby avoiding situations in which they have to put a last-second stop to their shower, which — analogous to finishing a task very close to a deadline — may be somewhat more stressful than it needed to be.

One implication of the second explanation is that the overall goal attainment rate should stay roughly constant or even rise over time, because the excess mass at the goal

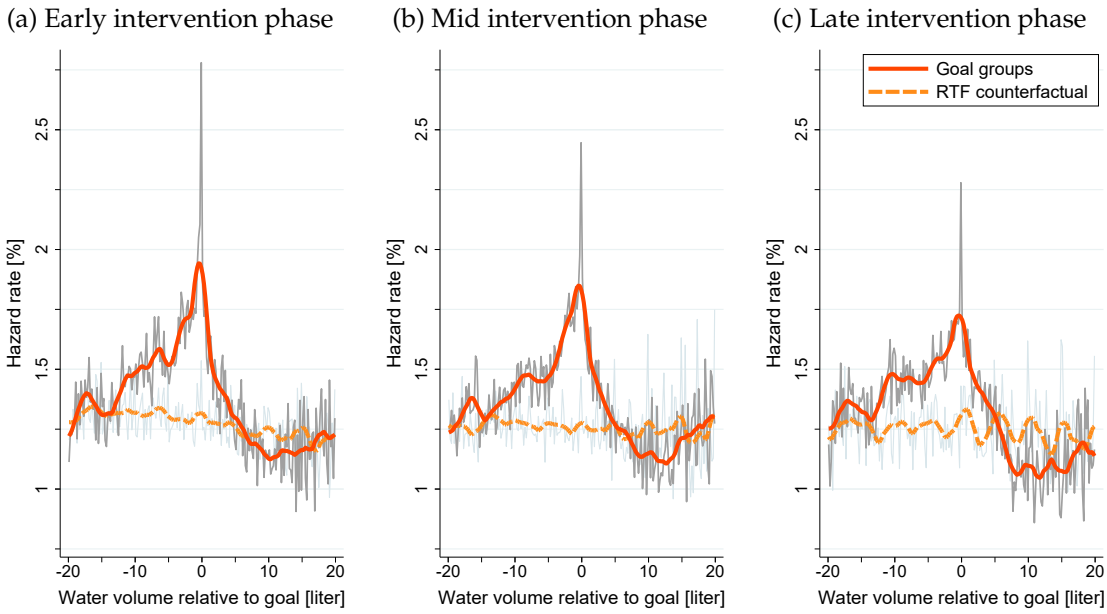


Figure 11: Stopping hazards over time

Notes: Hazard rates of showers by deciliter of distance to the respective conservation goal, split by three phases of the intervention period. Smoothed estimates are obtained through local linear regressions. All procedures follow the ones in Figure 10a.

would diminish simply by being diffused into lower water volume bins. In contrast, the first explanation would predict that the success rate decreases once individuals use it less as inflection point for evaluation. To distinguish between these two explanations, we therefore analyze whether and how goal attainment changes throughout the course of the study. Specifically, we estimate a linear probabilities model with a goal attainment dummy as outcome variable and study progress as regressor of interest — normalized such that its value is 0 at the start of the study and 1 after about four months. As benchmark, we look at hypothetical attainment rates by assigning placebo goals to households in the Control and RTF conditions using the same procedure as before.

The results are displayed in Table 7. In the baseline period, about 62% of showers would have met the conservation goal when pooling all difficulty levels. Columns (1) and (2) show that, hypothetically, attainment rate would have been higher for the RTF condition in the intervention period, which is simply due to the conservation effects in response to real-time feedback shown previously. As there is a slight general upward trend in water consumption levels in the months of our study (see Figure 7a), we observe corresponding decreases in hypothetical attainment rates from the beginning to the end of the intervention phase by about 1%p in the Control group and 1.5 %p in the RTF group, both statistically insignificant. When looking at actual attainment rates in the Goal conditions, column (3) shows that after an initial jump by 8% at the start of the intervention, the effect actually decreases by 3.8%p by the end of the four months study duration. This decrease is significantly larger than in the Control placebo ($p = 0.008$) and

Table 7: Overall goal attainment rates over time

	Placebo		*Actual attainment rates*	
	Control	RTF	Goal conditions (pooled)	
	(1)	(2)	(3)	(4)
Intervention	-0.009	0.017*	0.080***	0.021***
	(0.006)	(0.010)	(0.008)	(0.004)
Study progress	-0.010	-0.015	-0.038***	-0.011***
	(0.008)	(0.010)	(0.006)	(0.004)
Water volume FEs	–	–	–	*yes*
Bathroom FEs	*yes*	*yes*	*yes*	*yes*
Baseline mean	0.626	0.617	0.619	0.619
N	203275	181875	212680	212471
Clusters	70	67	360	360
R^2	0.175	0.189	0.348	0.715

Notes. Linear probabilties model. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

the RTF placebo ($p = 0.064$), suggesting that it is not only driven by broader trends in water consumption levels but also by goal-specific behavioral mechanisms. To further verify this, we additionally control for water volume fixed effects in column (4). If there was only one goal level in the sample, water volume would perfectly explain goal attainment in this specification. Thus, any non-zero coefficients can only be due to variation in the share of showers below a certain goal threshold relative to households who were assigned a different goal, e.g. when the likelihood of showers below 15 liters increases overproportionally in the 15L goal group.[26] The estimates show that, even conditional on water volume, goal success becomes significantly more likely once the intervention begins, again demonstrating that individuals respond specifically to the goal that was randomly assigned to them. Crucially, the conditional goal attainment effect drops by more than 50% by the end of the 4-month study period ($p = 0.004$).

Overall, the evidence here suggests that over the course of several weeks and months, individuals respond less to the goals that were assigned to them at the beginning of the study, as they gain a more nonchalant perspective on the feasibility and importance of missing or achieving that particular conservation goal. Nevertheless, we see in Section 4.3 that there is no evidence for a significant decease in average water conservation over the 4 month study period. Hence, externally-set goals seem to retain a status as vague norm or default about water consumption levels even though the precise target numbers associated with them become less psychologically binding.

[26]This is why in the Placebo checks using the RTF group and Control group, the coefficients and standard errors would precisely be 0 when adding water volume fixed effects.

6. Concluding remarks

In this paper, we presented evidence from a randomized field experiment in the context of household water conservation to examine the effectiveness of goal setting and its underlying behavioral mechanisms. Our experiment was designed to be representative of the population of Singapore and lasts between four to six months, which allows us to examine the long-term stability of goal setting as a behavioral policy tool. Importantly, our design allows us to cleanly separate the effects from providing neutral, quantitative feedback from the effect of goals. We further vary the difficulty of the goals by randomly assigning households to goal conditions ranging from 10L to 35L. Our results show that externally-set goals, when appropriately chosen, have a significant effect on conservation efforts. Among our five goal conditions, the 15L goal was the most effective in reducing water use, generating a treatment effect of 3.9 liters per shower, which is twice as high as the effect of real-time feedback alone. In line with the literature in psychology, the point estimates suggest that the best performing goals are challenging yet attainable. This does not only hold when comparing different groups, but also when analyzing heterogenous responses in different subgroups with regard to baseline water usage.

When analyzing fine-grained behavioral responses to goals, our data shows that the impact of goals on the stopping hazard of showers is particularly strong before individuals exceed the goal, with a large spike at the very large deciliter in which the goal is still achieved. In contrast, once individuals have missed the goal, the stopping hazard quickly decreases and becomes indistinguishable from the one in the experimental condition with only neutral feedback but not goals. Thus, while loss aversion in the form of higher marginal utility in the loss domain shapes behaviors in many domains (Sydnor, 2010; Fehr and Goette, 2007; Angrist et al., 2021), our evidence speaks against a prospect theory model of goals (Heath et al., 1999) and instead points toward a fixed psychological reward from achieving a goal, with little change in the marginal utility thereafter, as considered also by Allen et al. (2017). Thus, it may be more appropriate to interpret exogenous goals as norms or defaults for acceptable levels of water consumption. This is also supported by the fact that the easiest 35L goal seemed to be less effective than having no goal at all.

Interestingly, depending on the goal conditions, it can be the case that individuals repeatedly and consistently fail to meet their goal; vice versa, other individuals with (subjectively) easier goals may regularly achieve them without much effort. This raises the question of whether individuals stop paying attention to the goal over time, i.e. whether the goal effects are potentially short-lived. We find a very stark pattern in our average treatment effects: the full impact of the treatment materiallizes immediately, and remains stable over the entire study period of four to six months. There is no evidence of the effects vanishing over time as has been found with more aggregated forms of feedback (Houde et al., 2013). However, we document that the local responses to the specific goals

become significantly weaker over time, i.e. there is less bunching, and the goal attainment rate drops. These two seemingly contradictory observations, stable average effects and waning local effects, may be resolved by individuals forming habits or setting personal targets that replace the externally-set goal.

Overall, our study suggests that goal-setting (and real-time feedback) have the potential to be integrated into simple and easily scalable interventions to encourage desirable behavioral change for example in the domain of pro-environmental behavior, as modern digital technologies are becoming ever cheaper and more advanced.[27] Future research may also consider the comparison between the effectiveness of self-set goals and externally-set goals such as the ones we use in this study. Another important question is whether the effects of our interventions are limited only to that targeted activity, showering, or whether there are spillover effects to other water-consuming activities in the household. In a companion paper, we utilize billing data of households that participated in our experiment and observe statistically significant conservation effects of our interventions also in overall household water usage (Schmitt et al., 2021). Interestingly, the point estimates suggest quantitatively large *positive* spillover effects, i.e. reductions in water usage also outside of the shower, although we lack statistical power to detect spillover effects more precisely. An interesting avenue for further research is whether different types of interventions have an effect on the size and sign of spillover effects to non-targeted activities, as this may have important implications for cost-benefit calculations.

[27] The Public Utilities Board and the Housing Development Board have since launched an initiative to install smart shower meters in 10,000 newly built flats, with the configuration of the smart shower meter based on the 15L condition from this paper (PUB, 2018b).

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *NBER Working Paper 24003*, 2017.

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman, "Reference Points and Effort Provision," *American Economic Review*, 2011, *101* (2), 1–25.

Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter, "A review of intervention studies aimed at household energy conservation," *Journal of Environmental Psychology*, 2005, *25* (3), 273–291.

_ , _ , _ , and _ , "The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents," *Journal of Environmental Psychology*, 2007, *27* (4), 265–276.

Allcott, Hunt, "Social norms and energy conservation," *Journal of Public Economics*, 2011, *95* (9–10), 1082–1095.

_ and Sendhil Mullainathan, "Behavior and Energy Policy," *Science*, 2010, *327* (5970), 1204–1205.

_ and Todd Rogers, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, 2014, *104* (10), 3003–3037.

Allen, Eric J., Patricia M. Dechow, Devin G. Pope, and George Wu, "Reference-Dependent Preferences: Evidence from Marathon Runners," *Management Science*, 2017, *63* (6), 1657–1672.

Andor, Mark A. and Katja M. Fels, "Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects," *Ecological Economics*, 2018, *148*, 178–210.

Angrist, Joshua D., Sydnee Caldwell, and Jonathan V. Hall, "Uber versus Taxi: A Driver's Eye View," *American Economic Journal: Applied Economics*, 2021, *13* (3), 272–308.

Avery, Mallory, Osea Giuntella, and Peiran Jiao, "Why Don't We Sleep Enough? A Field Experiment Among College Students," *IZA Discussion Paper No. 12772*, 2019.

Ayres, I., S. Raseman, and A. Shih, "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage," *Journal of Law, Economics, and Organization*, 2013, *29* (5), 992–1022.

Blanes i Vidal, Jordi and Mareike Nossol, "Tournaments Without Prizes: Evidence from Personnel Records," *Management Science*, 2011, *57* (10), 1721–1736.

Brookins, Philip, Sebastian J. Goerg, and Sebastian Kube, "Self-chosen goals, incentives, and effort," *Working Paper*, 2017.

Byrne, David, Lorenz Goette, Leslie A. Martin, Lucy Delahey, Alana Jones, Amy Miles, Samuel Schoeb, Thorsten Staake, and Verena Tiefenbeck, "The Habit-Forming Effects of Feedback: Evidence From a Large-Scale Field Experiment," *CRC TR 224 Discussion Paper No. 285*, 2021.

Camerer, Colin, Linda Babcock, George Loewenstein, and George Thaler, "Labor Supply of New York City Cabdrivers: One Day at a Time," *Quarterly Journal of Economics*, 1997, *112* (2), 407–441.

Carlsson, Fredrik, Christina Annette Gravert, Verena Kurz, and Olof Johansson-Stenman, "The Use of Green Nudges as an Environmental Policy Instrument," *Review of Environmental Economics and Policy*, 2021, *15* (2), 216–237.

Cerulli-Harms, Annette, Lorenz Goette, and Charles Sprenger, "Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences," *American Economic Journal: Microeconomics*, 2019, *11* (1), 185–207.

Chapman, Gretchen B., Helen Colby, Kimberly Convery, and Elliot J. Coups, "Goals and Social Comparisons Promote Walking Behavior," *Medical Decision Making*, 2015, *advance online publication. doi: 10.1177/0272989X15592156.*

Charness, Gary and Uri Gneezy, "Incentives to Exercise," *Econometrica*, 2009, *77* (3), 909–931.

Chetty, Raj, John N. Friedman, Tore Olsen, and Luigi Pistaferri, "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records," *Quarterly Journal of Economics*, 2011, *126* (2), 749–804.

Clark, Damon, David Gill, Victoria Prowse, and Mark Rush, "Using Goals to Motivate College Students: Theory and Evidence From Field Experiments," *Review of Economics and Statistics*, 2020, *102* (4), 648–663.

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-González, "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough," *Management Science*, 2015, *61* (12), 2926–2944.

Crawford, Vincent P. and Juanjuan Meng, "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 2011, *101* (5), 1912–1932.

Creutzig, Felix, Joyashree Roy, William F. Lamb, Inês M. L. Azevedo, Wändi Bruine de Bruin, Holger Dalkmann, Oreane Y. Edelenbosch, Frank W. Geels, Arnulf Grubler, Cameron Hepburn, Edgar G. Hertwich, Radhika Khosla, Linus Mattauch, Jan C. Minx, Anjali Ramakrishnan, Narasimha D. Rao, Julia K. Steinberger, Massimo Tavoni, Diana Ürge-Vorsatz, and Elke U. Weber, "Towards demand-side solutions for mitigating climate change," *Nature Climate Change*, 2018, *8* (4), 260–263.

Della Vigna, Stefano, "Psychology and Economics : Evidence from the Field," *Journal of Economic Literature*, 2009, *472*, 315–372.

DellaVigna, Stefano and Ulrike Malmendier, "Paying Not to Go to the Gym," *American Economic Review*, 2006, *96* (3), 694–719.

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio, "Information Strategies and Energy Conservation Behavior: A Meta-analysis of Experimental Studies from 1975 to 2012," *Energy Policy*, 2013, *61*, 729–739.

Diecidue, Enrico and Jeroen van de Ven, "Aspiration Level, Probability of Success and Failure, and Expected Utility," *SSRN Electronic Journal*, 2006.

Dietz, Thomas, Gerald T. Gardner, Jonathan Gilligan, Paul C. Stern, and Michael P. Vandenbergh, "Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions," *Proceedings of the National Academy of Sciences*, 2009, *106* (44), 18452–18456.

Dobronyi, Christopher R., Philip Oreopoulos, and Uros Petronijevic, "Goal Setting, Academic Reminders, and College Success: A Large-Scale Field Experiment," *Journal of Research on Educational Effectiveness*, 2019, *12* (1).

Dohmen, Thomas, Armin Falk, Klaus Fliessbach, Uwe Sunde, and Bernd Weber, "Relative versus absolute income, joy of winning, and gender: Brain imaging evidence," *Journal of Public Economics*, 2011, *95* (3-4), 279–285.

Drexler, Alejandro, Greg Fischer, and Antoinette Schoar, "Keeping It Simple: Financial Literacy and Rules of Thumb," *American Economic Journal: Applied Economics*, 2014, *6* (2), 1–31.

Drucker, Peter F., *The Practice of Management*, New York: Harper & Row, 1954.

Edwards, E. A., J. Lumsden, C. Rivas, L. Steed, L. A. Edwards, A. Thiyagarajan, R. Sohanpal, H. Caton, C. J. Griffiths, M. R. Munafò, S. Taylor, and R. T. Walton, "Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps," *BMJ Open*, 2016, *6* (10).

Ericson, Keith and Andreas Fuster, "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments," *Quarterly Journal of Economics*, 2011, *forthcomin.*

Fan, James and Joaquín Gómez-Miñambres, "Nonbinding Goals in Teams: A Real Effort Coordination Experiment," *Manufacturing & Service Operations Management*, 2020, *22* (5), 1026–1044.

Fang, Ximeng, Lorenz Goette, Bettina Rockenbach, Matthias Sutter, Verena Tiefenbeck, Samuel Schoeb, and Thorsten Staake, "Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Energy Conservation," *CRC TR 224 Discussion Paper No. 149*, 2020.

Farber, Henry S., "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers," *Journal of Political Economy*, 2005, *113* (1), 46–82.

_ , "Why you Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers*," *Quarterly Journal of Economics*, 2015, *130* (4), 1975–2026.

Fehr, Ernst and Lorenz Goette, "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment," *American Economic Review*, 2007, *97* (1), 298–317.

Ferraro, Paul J. and Juan José Miranda, "Heterogeneous treatment effects and mechanisms in information-based environmental policies: Evidence from a large-scale field experiment," *Resource and Energy Economics*, 2013, *35* (3), 356–379.

_ **and Michael K. Price**, "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment," *Review of Economics and Statistics*, 2013, *95* (1), 64–73.

Fischer, Corinna, "Feedback on Household Electricity Consumption: A Tool for Saving Energy?," *Energy Efficiency*, 2008, *1* (1), 79–104.

Frederiks, Elisha R., Karen Stenner, and Elizabeth V. Hobman, "Household Energy Use: Applying Behavioural Economics to Understand Consumer Decision-Making and Behaviour," *Renewable and Sustainable Energy Reviews*, 2015, *41*, 1385–1394.

Gallus, Jana, "Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia," *Management Science*, 2017, *63* (12), 3999–4015.

Gerster, Andreas, Mark Andor, and Lorenz Goette, "Disaggregate Consumption Feedback and Energy Conservation," *CEPR Discussion Paper 14952*, 2020.

Gneezy, Uri, Lorenz Goette, Charles Sprenger, and Florian Zimmermann, "The limits of expectations-based reference dependence," *Journal of the European Economic Association*, 2017, *15* (4), 861–876.

Goette, Lorenz, Hua-Jing Han, and Zhi Hao Lim, "The Dynamics of Goal Setting: Evidence From a Field Experiment on Resource Conservation," *CRC TR 224 Discussion Paper No. 283*, 2021.

_ , Thomas Graeber, Alexandre Kellogg, and Charles Sprenger, "Heterogeneity of Loss Aversion and Expectations-Based Reference Points," *Working Paper*, 2020.

Gómez-Miñambres, Joaquín, "Motivation through goal setting," *Journal of Economic Psychology*, 2012, *33* (6), 1223–1239.

Harding, Matthew and Alice Hsiaw, "Goal setting and energy conservation," *Journal of Economic Behavior & Organization*, 2014, *107*, 209–227.

Heath, Chip, Richard Larrick, and George Wu, "Goals as Reference Points," *Cognitive Psychology*, 1999, *38*, 79–107.

Höpfner, Jessica and Nina Keith, "Goal Missed, Self Hit: Goal-Setting, Goal-Failure, and Their Affective, Motivational, and Behavioral Consequences," *Frontiers in psychology*, 2021, *12*, 704790.

Houde, Sebastien, Annika Todd, Anant Sudarshan, June A. Flora, and K. Carrie Armel, "Real-time Feedback and Electricity Consumption: A Field Experiment Assessing the Potential for Savings and Persistence," *The Energy Journal*, 2013, *34* (1).

Kahneman, Daniel and Amos Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, 1979, *47* (2), 263–291.

Karlin, Beth, Joanne F. Zinger, and Rebecca Ford, "The effects of feedback on energy conservation: A meta-analysis," *Psychological bulletin*, 2015, *141* (6), 1205–1227.

Kleven, Henrik J. and Mazhar Waseem, "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," *Quarterly Journal of Economics*, 2013, *128* (2), 669–723.

Kleven, Henrik Jacobsen, "Bunching," *Annual Review of Economics*, 2016, *8* (1), 435–464.

Koch, Alexander K. and Julia Nafziger, "Self-regulation through Goal Setting," *Scandinavian Journal of Economics*, 2011, *113* (1), 212–227.

_ and _ , "Goals and bracketing under mental accounting," *Journal of Economic Theory*, 2016, *162*, 305–351.

Kollmuss, Anja and Julian Agyeman, "Mind the Gap: Why Do People Act Environmentally and What Are the Barriers to Pro-Environmental Behavior?," *Environmental Education Research*, 2002, *8* (3), 239–260.

Kosfeld, Michael and Susanne Neckermann, "Getting More Work for Nothing? Symbolic Awards and Worker Performance," *American Economic Journal: Microeconomics*, 2011, *3* (3), 86–99.

Koszegi, Botond and Matthew Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 2006, *121* (4), 1133–1165.

_ **and** _ , "Reference-Dependent Consumption Plans," *American Economic Review*, 2009, *99* (3), 909–936.

Kuhn, Peter and Lizi Yu, "Kinks as Goals: Accelerating Commissions and the Performance of Sales Teams," *IZA Discussion Paper No. 14115*, 2021.

Locke, Edwin A. and Gary P. Latham, *A Theory of Goal Setting and Task Performance*, Englewood Cliffs, NJ: Prentice-Hall, 1990.

_ **and** _ , "Building a practically useful theory of goal setting and task motivation. A 35-year odyssey," *The American psychologist*, 2002, *57* (9), 705–717.

_ **and** _ , *New developments in goal setting and task performance*, New York: Routledge, 2013.

_ **and** _ , "Does prospect theory add or subtract from our understanding of goal directed motivation?," in Diana L. Stone and James H. Dublebohn, eds., *The Only Constant in HRM Today is Change*, Charlotte, NC: Information Age Publishing, 2019, pp. 19–42.

_ **and** _ , "The development of goal setting theory: A half century retrospective," *Motivation Science*, 2019, *5* (2), 93–105.

Loock, Claire-Michelle, Thorsten Staake, and Frédéric Thiesse, "Motivating Energy-Efficient Behavior With Green IS: An Investigation of Goal Setting and the Role of Defaults," *MIS Quarterly*, 2013, *37* (4), 1313–1332.

Madrian, Brigitte C., "Applying Insights From Behavioral Economics To Policy Design," *Annual Review of Economics*, 2014, *6*, 663–688.

Markle, Alex, George Wu, Rebecca White, and Aaron Sackett, "Goals as reference points in marathon running: A novel test of reference dependence," *Journal of Risk and Uncertainty*, 2018, *56* (1), 19–50.

Mento, Anthony J., Robert P. Steel, and Karren, Ronald, J., "A Meta-Analytic Study of the Effects of Goal-Setting on Task Performance: 1996-1984," *Organizational Behavior and Human Decision Processes*, 1987, *39*, 52–83.

PUB, "Singapore World Water Day 2018 & Household Water Consumption Study," *Press Release by Singapore's National Water Agency*, 2018, *March 1.*

_, "Smart Shower Programme," *https://www.pub.gov.sg/savewater/athome/smartshowerprogramme (accessed June 7, 2018)*, 2018.

Reddy, Sheila M.W., Jensen Montambault, Yuta J. Masuda, Elizabeth Keenan, William Butler, Jonathan R.B. Fisher, Stanley T. Asah, and Ayelet Gneezy, "Advancing Conservation by Understanding and Influencing Human Behavior," *Conservation Letters*, 2017, *10* (2), 248–256.

Samek, Anya, "Gifts and goals: Behavioral nudges to improve child food choice at school," *Journal of Economic Behavior & Organization*, 2019, *164*, 1–12.

Schmitt, Kathrin, Verena Tiefenbeck, Ximeng Fang, Lorenz Goette, Thorsten Staake, and Davin Wang, "Pro-environmental spillover effects in the resource conservation domain: Evidence from a randomized controlled trial in Singapore," *mimeo*, 2021.

Sydnor, Justin, "(Over)insuring Modest Risks," *American Economic Journal: Applied Economics*, 2010, *2* (4), 177–199.

Taylor, Michael and Claudio Accheri, "This is Singapore's plan to avoid running out of water," *World Economic Forum*, 2019, *August 13 (https://www.weforum.org/agenda/2019/08/singapore-focus-innovation-key-securing-water-future).*

Thakral, Neil and Linh T. Tô, "Daily Labor Supply and Adaptive Reference Points," *American Economic Review*, 2021, *111* (8), 2417–2443.

Thaler, Richard H. and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008.

Tiefenbeck, Verena, Anselma Woerner, Samuel Schoeb, Elgar Fleisch, and Thorsten Staake, "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives," *Nature Energy*, 2019, *4*, 35–41.

_, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake, "Overcoming salience bias: how real-time feedback fosters resource conservation," *Management Science*, 2018, *64* (3), 1458–1476.

Wood, Wendy and Dennis Rünger, "Psychology of Habit," *Annual review of psychology*, 2016, *67*, 289–314.

For Online Publication

Appendix A Supplementary figures and tables

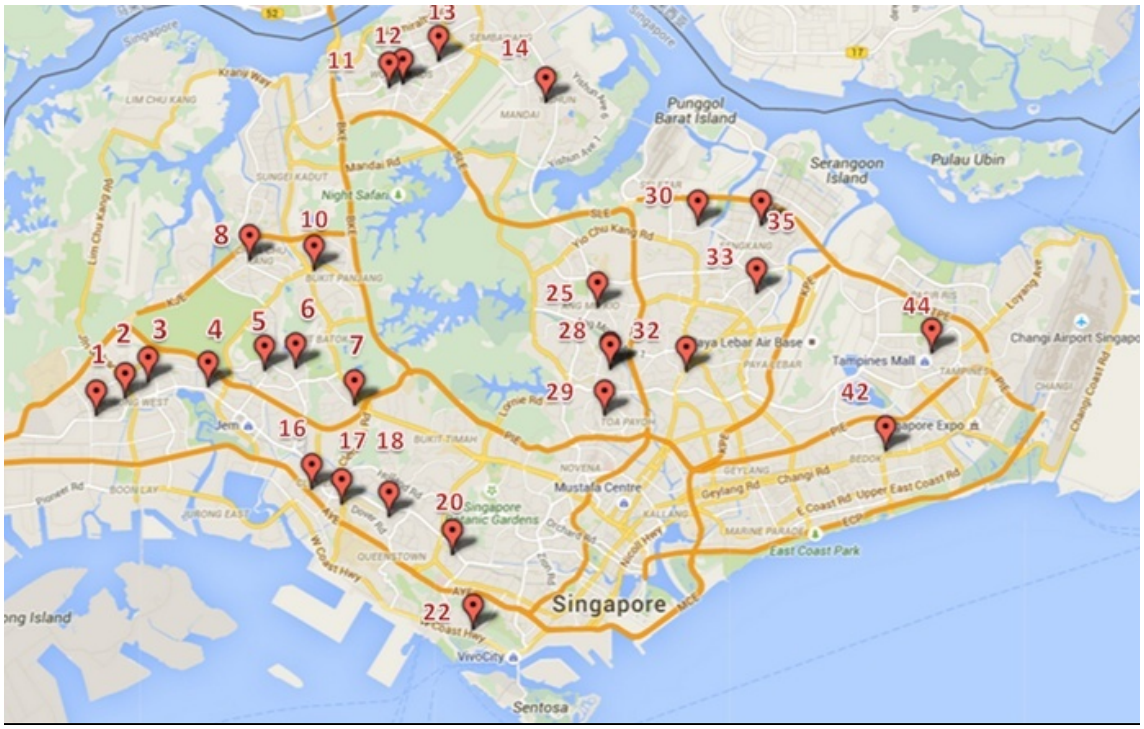


Figure A1: Sites of participating households in Singapore.

Table A1: Baseline shower information – overview

	Average	SD	25th pctile	Median	75th pctile	Observationss
Volume [liter]	20.03	16.66	8.90	14.90	25.30	15500
Flow rate [l/min]	5.26	2.35	3.60	4.60	6.40	15500
Temperature [Celsius]	33.77	3.01	31	34	36	15460
Duration [min]	4.91	3.74	2.45	3.87	6.18	15500

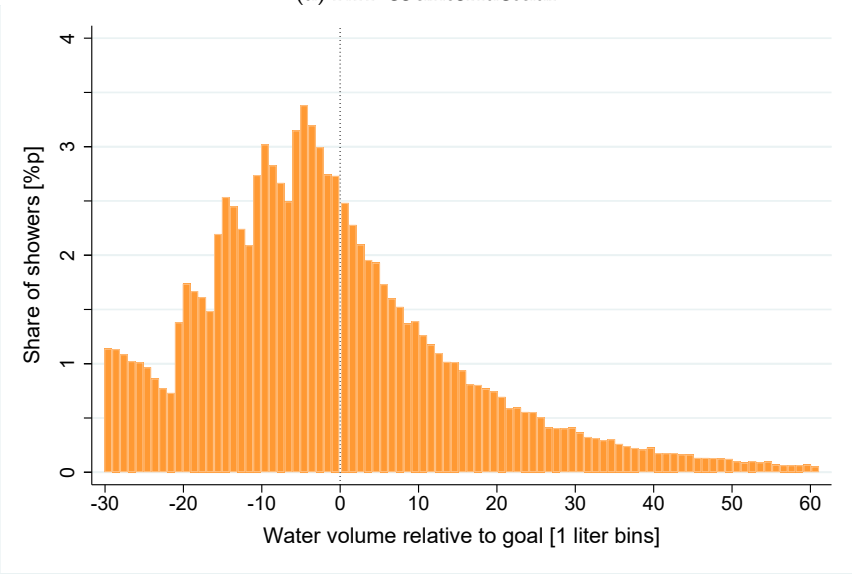
Notes: 775 devices with at least 20 showers and valid data records are considered. For water temperature statistics, 2 devices with broken temperature sensors are excluded. The shower duration only considers time with water flow, i.e. excluding breaks.

Table A2: Treatment effect on number of showers

	(1) Total	(2) Total	(3) Total	(4) Person-Day
10 liter goal	-21.30	-7.39	-12.03	0.04
	(37.14)	(39.60)	(34.82)	(0.09)
15 liter goal	-0.41	-2.64	14.05	0.05
	(37.37)	(39.74)	(37.46)	(0.09)
20 liter goal	21.52	-2.38	-7.81	0.11
	(37.06)	(40.87)	(36.91)	(0.09)
25 liter goal	-10.93	-17.39	22.51	0.14
	(37.29)	(36.57)	(34.35)	(0.09)
35 liter goal	12.48	12.82	41.91	0.15
	(37.37)	(39.49)	(38.36)	(0.10)
Real-time feedback	-8.96	-0.57	12.12	0.08
	(37.97)	(42.09)	(38.95)	(0.10)
Constant	390.48***	423.48***	409.14***	1.19***
	(26.31)	(29.76)	(27.74)	(0.07)
[Controls]	No	No	Yes	No
Devices with fewer than 40 showers	Yes	No	No	No
Observations	822	747	707	442
R^2	0.002	0.001	0.202	0.009
$\beta_{10L} = \ldots = \beta_{35L} = \beta_{RT} = 0$	$p = 0.93$	$p = 0.99$	$p = 0.73$	$p = 0.67$

Control variables include the time between deployment and retrieval, number of adults and children in the household, and interactions of both. In columns (3) and (4), devices sent back via postal service are excluded. In column (4), households with study duration shorter than 3 months and top and bottom percentiles are cut off. Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(a) RTF counterfactual



(b) Goal conditions

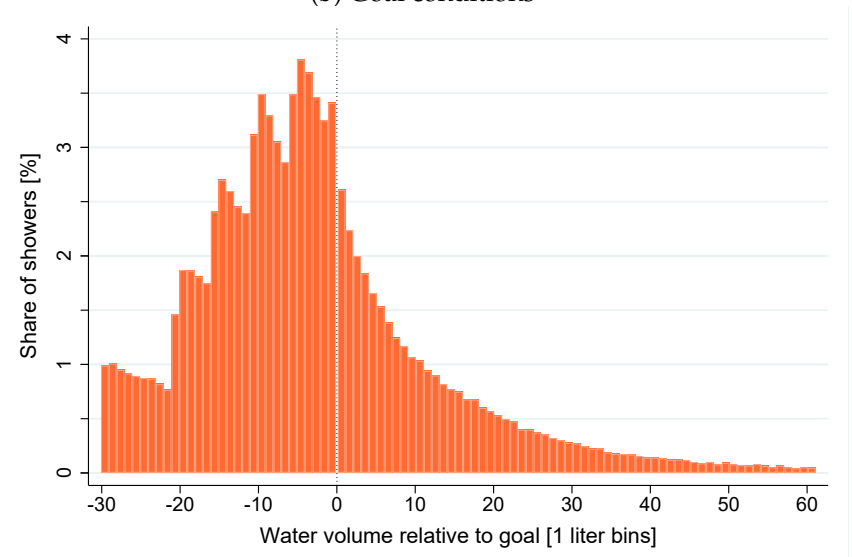


Figure A2: Distribution of intervention period showers

Notes: Awesome description.

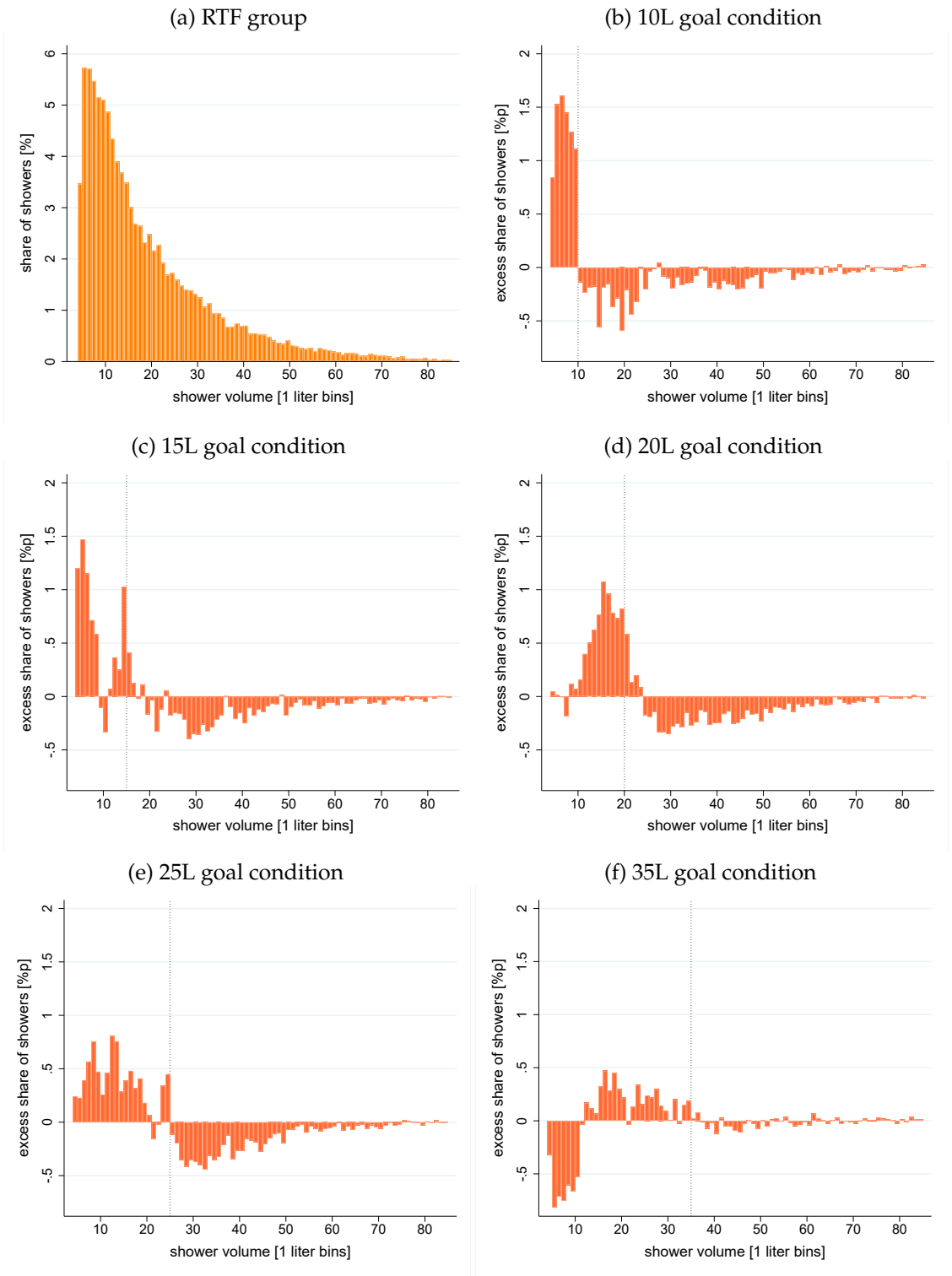


Figure A4: Excess mass of goal groups relative to RTF group

Notes: Bla.

Table A3: Randomization checks — water conservation attitudes

	Try generally to ...		*Conserve water to ...*	
	protect the environment	save money	protect the environment	save money
RTF group	-0.101	-0.043	-0.077	0.126
	(0.132)	(0.154)	(0.098)	(0.122)
10L goal group	0.071	0.014	0.086	0.086
	(0.132)	(0.162)	(0.091)	(0.133)
15L goal group	0.032	0.046	-0.004	0.111
	(0.131)	(0.180)	(0.101)	(0.116)
20L goal group	-0.076	0.022	-0.091	0.163
	(0.138)	(0.168)	(0.108)	(0.122)
25L goal group	-0.058	-0.002	-0.052	-0.033
	(0.133)	(0.156)	(0.090)	(0.135)
35L goal group	-0.090	0.105	0.020	0.163
	(0.128)	(0.181)	(0.097)	(0.116)
Constant	0.743***	-0.286***	1.271***	1.143***
	(0.095)	(0.108)	(0.067)	(0.096)
Observations	495	495	495	495
R^2	0.006	0.002	0.009	0.011
p-value of joint null	0.787	0.993	0.547	0.566

Only includes households that are included in the main analysis sample. Missing responses for four households in these survey questions. Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

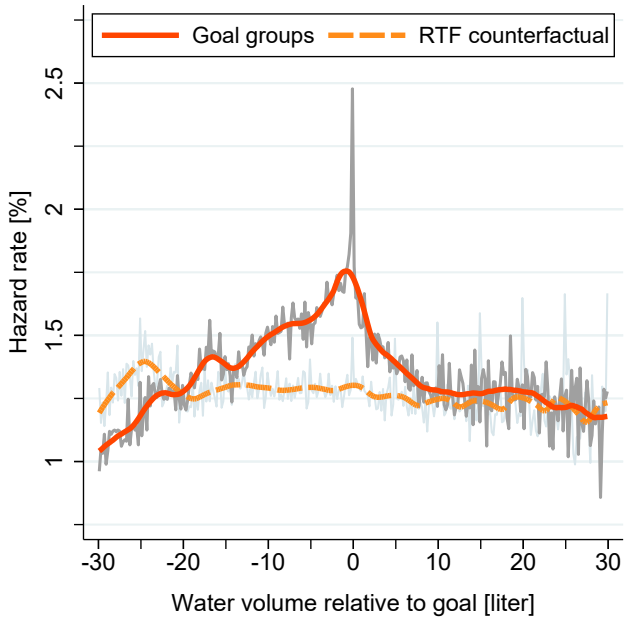


Figure A5: Stopping hazard by goal distance — not adjusted for baseline differences.

Notes: Bla

Table A4: Stability of treatment effects: four-part splines with study progress

		× progress splines			
	initial effect	1st spline	2nd spline	3rd spline	4th spline
10l goal × intervention	-3.232***	0.009	0.004	0.003	0.006
	(0.593)	(0.017)	(0.013)	(0.017)	(0.018)
15l goal × intervention	-3.974***	0.012	0.008	-0.013	0.020
	(0.662)	(0.016)	(0.013)	(0.015)	(0.018)
20l goal × intervention	-2.956***	-0.003	0.016	-0.021	0.031
	(0.560)	(0.015)	(0.013)	(0.016)	(0.024)
25l goal × intervention	-2.815***	-0.010	0.010	0.005	0.018
	(0.565)	(0.015)	(0.012)	(0.016)	(0.020)
35l goal × intervention	-1.938***	0.025	0.003	-0.012	-0.006
	(0.556)	(0.018)	(0.014)	(0.017)	(0.020)
Real-time feedback × intervention	-1.558***	-0.010	0.012	-0.014	0.005
	(0.552)	(0.017)	(0.014)	(0.016)	(0.023)
Constant	19.668***				
	(0.237)				
F-test: all 10l goal splines = 0			$p = 0.9464$		
F-test: all 15l goal splines = 0			$p = 0.5287$		
F-test: all 20l goal splines = 0			$p = 0.4848$		
F-test: all 25l goal splines = 0			$p = 0.7281$		
F-test: all 35l goal splines = 0			$p = 0.5934$		
F-test: all RTF splines = 0			$p = 0.8419$		
F-test: all splines = 0			$p = 0.7268$		
Observations			313996		
R^2			0.332		

1st progress spline defined from 6 to 37, 2nd progress spline defined from 37 to 68, 3rd spline defined from 69 to 100, 4th spline defined from 101 to 150 (6 month devices). Standard errors in parentheses (clustered on household level). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Probability of showers just above or below a salient threshold

	below salient threshold			above salient threshold		
	0.5L bin	1L bin	2L bin	0.5L bin	1L bin	2L bin
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.008	0.002	0.014	0.005	0.007	-0.022
	(0.010)	(0.012)	(0.015)	(0.010)	(0.011)	(0.014)
Treated × goal group	-0.010	-0.006	-0.014	-0.006	-0.011	0.008
	(0.008)	(0.010)	(0.012)	(0.008)	(0.008)	(0.011)
Matching goal	0.042***	0.046***	0.039***	0.002	-0.007	-0.032***
	(0.007)	(0.007)	(0.008)	(0.004)	(0.005)	(0.008)
Matching goal × study progress	-0.029***	-0.028***	-0.009	-0.003	0.002	0.012
	(0.008)	(0.009)	(0.011)	(0.006)	(0.008)	(0.011)
Intervention	0.013**	-0.000	0.001	0.001	0.002	0.012
	(0.006)	(0.008)	(0.010)	(0.006)	(0.008)	(0.010)
Constant	0.095***	0.159***	0.322***	0.088***	0.145***	0.282***
	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)
Bathroom fixed effects	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Threshold fixed effects	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
N	289710	289710	289710	289710	289710	289710
R^2	0.039	0.064	0.158	0.033	0.057	0.128

Notes. Results come from estimating equation 3 using ordinary least squars. The dependent variable is an indicator for whether a shower falls into a particular volume bin around a salient threshold. We consider thresholds in steps of 5 from 10 liters to 45 liters. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$