The Dynamics of Goal-Setting: Evidence from a Natural Field Experiment*

Lorenz Goette†, Hua-Jing Han‡, Zhi Hao Lim§

June 4, 2025

Abstract

In a field experiment across two residential colleges, we investigate responses to evolving goal difficulty using real-time feedback and moral suasion in resource conservation. In phase 1, both moderate and hard goals lead to similar conservation effects. In phase 2, after adjusting goals to an intermediate level, treatment effects diverge due to the hard-goal group's underperformance, reducing the effect by 30%. Throughout the intervention, the moderate-goal group's high baseline users maintain stronger conservation effects than their counterparts in the hard-goal group. Our findings suggest excessively challenging goals can damage motivation, with subsequent goal adjustments failing to reverse the initial imprint.

JEL classification: C93, Q41, D91

Keywords: field experiment, goal-setting, resource conservation

*We are grateful to the NUS Office of Housing Services and our research assistant team in Singapore for their invaluable support in implementing the experiment. We thank Ximeng Fang, Sebastian Kube, and Johannes Weber for helpful comments. This research was supported by the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B07) and ECONtribute (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2126/1- 390838866. All errors that remain are ours.

†Goette: National University of Singapore (email: ecslfg@nus.edu.sg).

‡Han: University of Bonn (email: han@uni-bonn.de).

§Lim: Columbia University (email: zl2969@columbia.edu).

1 Introduction

Goals are pivotal in motivating individuals and enhancing task performance. Across both private and public sectors, they are considered a key instrument for managing employee motivation and effort provision (see, e.g. Drucker, 1954; Grove, 1983). In economics, research has delved into how contracts that feature discrete goals, coupled with bonus incentives, can serve as optimal incentive schemes for addressing moral hazard problems.[1] Additionally, a large literature in psychology supports the view that goals are important determinants of task performance (Locke, 1968; Locke et al., 1981), and they can have effects above and beyond the financial incentives associated with them (Pritchard and Curts, 1973; Asmus et al., 2015).

Moreover, goals are not static and can change over time. In management, business objectives often evolve in response to shifts in the economic environment or the internal development of a firm (see, e.g. Kennerley and Neely, 2003; Fisher et al., 2016). If goals directly affect individuals' motivation and performance, changing them can pose additional challenges. In particular, the way goals are initially set and subsequently adjusted can influence individuals' effort in reaching those goals.

Consider a scenario where two (otherwise identical) individuals start with different goals: one with a moderate goal M and the other with a hard goal H. If goals act as reference points (Heath et al., 1999), individuals may exert more effort to reach M, while diminishing sensitivity leads them to exert less effort to try and reach H.[2] Now suppose both individuals are reassigned to a common intermediate goal I, which is halfway between M and H. For the individual who started at M, the goal has become harder. However, because their reference point remains near M, loss aversion incentivizes them to work harder in order to try and reach I. By contrast, for the individual who began with the hard goal H, the new goal now feels easier relative to their reference point, placing them in the gain domain and reducing incentive to exert effort. Consequently, even though both individuals now face the same goal I, those who started out with the moderate goal M may perform better in the long run than those who started out with

[1] For instance, Oyer (2000) shows how these schemes can be optimal under limited-liability constraints, while Levin (2003) examines how such contracts can emerge as efficient equilibria in relational contracts within repeated-game settings. More generally, Abreu et al. (1990) show that 'bang-bang equilibria' - akin to a goal with a bonus - can arise in repeated games with imperfect monitoring.

[2] Heath et al. (1999) show in a series of hypothetical choices that goals inherit the properties of reference points, and individuals respond in a manner consistent with experiencing loss aversion and diminishing sensitivity around them as described in prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991). This view is also consistent with the Kőszegi and Rabin (2006) model of reference-dependent preferences, in which reference points are given by recent expectations: goals may affect expectations, and thus create a reference point.

the hard goal H.

This paper presents a large-scale field experiment on shower water conservation to test this hypothesis. We examine how the difficulty of initial goals affects effort and performance, and whether subsequently adjusting goals to a common intermediate level reinforces or undermines previously established conservation efforts.

Our experiment involved around $1,000$ student residents from two residential colleges at the National University of Singapore (NUS). We installed smart shower heads in all designated bathrooms to enable real-time data collection on shower water usage. To study the effects of changing goals on water conservation, we employed two key interventions: moral suasion and real-time feedback. Notably, our interventions were embedded within a broader university-led resource conservation initiative; thus, the residents were unaware that they were part of an experiment, qualifying it as a natural field experiment (Harrison and List, 2004). Another aspect of our study is that the residents paid a flat fee for their monthly rent and were not charged for water and energy use. Taken together, these features allow us to effectively rule out self-selection bias and isolate the effects of goal-setting on non-pecuniary motivations.

Our intervention comprised two phases. In phase 1, we assigned subjects to one of four experimental groups. In our two key treatment groups, subjects were encouraged to keep their water use in showers below a specific target: one group received a challenging 18L goal, and the other received a more achievable 28L goal, considering the baseline average water usage of around 32L. The 18L goal, requiring almost a 50% reduction in water use, was designated as the hard goal, while the 28L goal, requiring a reduction of about 15%, was considered moderate. To assist subjects in meeting these targets, the smart shower heads were programmed to provide real-time feedback on water usage through color changes relative to the assigned goal.[3] This feedback mechanism was communicated through informative posters in the shower facilities. We label these two treatment groups as the 18L goal and 28L goal groups, respectively. A third experimental group, aimed at controlling for moral suasion effects, received only the conservation message through posters, encouraging them to keep their water use below 28L, but without real-time feedback on their shower use. Finally, the control group received neither real-time feedback nor moral suasion. In phase 2, we maintained the original setup but adjusted the initial goals of either 18L or 28L to the same intermediate goal of 24L, which is halfway between the two initial

[3] At the beginning of a shower, the smart shower head emits a green light. As water usage increases, the light changes color sequentially: first to yellow, then to orange, and finally to red. When the water volume exceeds the assigned goal, the shower head signals this by displaying a blinking red light.

goals. Similarly, for the moral suasion group, we assigned the same 24L goal, communicated through updated posters.

Our findings reveal the nuanced impact of changing goal difficulty on conservation behavior. We document that moral suasion alone did not significantly affect shower water use in either phase of the experiment. In contrast, combining goal-setting with real-time feedback led to large and significant reductions in water use, averaging around a 15% decrease from the baseline. Interestingly, in phase 1, there was no significant difference in the average treatment effects (ATEs) between the 18L goal and 28L goal groups. However, when the goal was reset to 24L for both groups in phase 2, a notable divergence in ATEs emerged: subjects initially assigned the hard goal of 18L systematically underperformed compared to their counterparts who began with the moderate goal of 28L. The differences are quantitatively important; the group that started with the hard goal exhibited a 30% lower ATE in phase 2 compared to the group that began with the moderate goal.

These treatment effects were especially pronounced among high baseline users. Those initially assigned the hard goal of 18L displayed a 45% lower ATE in phase 2 compared to their counterparts in the 28L goal group. We dig deeper into the role of this heterogeneity using the causal forest method (Athey et al., 2019; Athey and Wager, 2019). The results offer a more clear-cut interpretation of where the heterogeneity in treatment effects comes from. In phase 1, within the 18L goal group, we show that as baseline consumption increased, individuals gave up earlier in trying to meet the goal. In contrast, in the 28L goal group, individuals with the same baseline consumption exerted more effort to reduce water use. When the goal was reset to 24L for everybody in phase 2, we observe two counteracting forces in the 18L goal group: those with a baseline consumption below 30L increased their conservation efforts relative to phase 1, whereas individuals with baseline usage above 35L reduced their efforts further. By contrast, tightening the goal to 24L for the 28L goal group led to increased conservation efforts among those with baseline consumption around 30L, and only a slight dropoff among the very high baseline users, as the harder 24L goal may have become less attainable for them. Taken together, these results suggest a special role of assigning the initial goal. Specifically, they indicate that individuals who initially disengaged due to high baseline consumption did not re-engage when the goal was adjusted to an intermediate level. This shows that the initial imprint of a goal is not easily undone by simply resetting it, highlighting the lasting influence of initial goal assignments on behavior.

Our work is related to two main strands of literature. First, it contributes to the literature on the role of goal-setting in shaping motivation and effort provision. Across a range of domains, including workplace behavior, higher education, and energy conservation, studies show that both externally assigned and self-set goals can enhance effort and task performance, particularly when the goals are perceived as attainable (Corgnet et al., 2015; Asmus et al., 2015; Clark et al., 2020; Brookins et al., 2017; Harding and Hsiaw, 2014). Even an individual's personal best can serve as a specific and difficult goal to increase effort, as evidenced in online chess games (Anderson and Green, 2018).[4] In their widely acknowledged paper, Locke and Latham (2002) describe a positive linear relationship between goal difficulty and task performance, assuming individuals accept and are committed to the goal. In particular, Erez (1977) emphasizes the importance of feedback as a necessary condition for a positive relationship between goals and performance. However, it has been observed that performance may level out or even decline when goals become highly difficult and are rejected (Erez and Zidon, 1984). Notwithstanding, the literature has not explored the potential side effects on performance associated with changing goals over time, which our paper addresses.[5] In line with the findings of Goerg et al. (2019) and Agarwal et al. (2018), we show that harder goals can lead to worse outcomes. However, we additionally demonstrate that these demotivating effects carry over to the next phase, even when both goals are adjusted to the same level of difficulty. To our knowledge, we are the first to show that setting too hard a goal can lead to a lasting, detrimental imprint, reducing effort and performance, even after the goal had been adjusted to an intermediate level. The initial goal assignment is thus critical and has long-lasting effects on task performance, which cannot be easily undone. In our setting, this effect was particularly pronounced amongst high baseline individuals: the hard goal muted their conservation efforts right from the outset, and this effect persisted throughout the intervention.

Second, our paper speaks to a rich literature on behavioral interventions for resource conservation, contributing new evidence on how the dynamics of goal-setting influence sustained

[4] In related work, Avoyan et al. (2020) document that some online chess players are more likely to stop after a win (called *win-stoppers*), while others are more likely to stop after a loss (called *loss-stoppers*); their finding suggests that accounting for preference heterogeneity may have implications for the design of goals to achieve various objectives. Drawing parallels from another domain, Haenni (2019) find that in amateur tennis tournaments, players tend to delay their participation in the next tournament following a loss. In particular, the impact of losing against a lower ranked opponent had a significantly larger effect on the delay compared to losing against a higher ranked opponent, thus suggesting that a player's ranking may serve as a reference point for evaluating actual defeats.

[5] Latham and Locke (2006) list ten possible pitfalls of goal-setting which include conflict among group members, cheating for monetary incentives, perceiving the goal as a threat, etc. Ordóñez et al. (2009) further argue the harmful effects of goal-setting such as distorting risk preferences, promoting unethical behavior, and diminishing intrinsic motivation. But none of these specifically addresses the potential pitfalls that might arise when goals change over time, except for the fact that in changing business environments, performance goals might actually prevent learning.

behavioral change. Recent research highlights that limited attention and imperfect information by households play an important role in shaping resource consumption (Chetty et al., 2009; Attari et al., 2010; Tiefenbeck et al., 2018; Langenbach et al., 2019). One promising intervention is to supplement goal-setting with the provision of information feedback; this has been shown to significantly reduce resource consumption (see, e.g. Harding and Hsiaw, 2014; Becker, 1978; Abrahamse et al., 2005; Attari et al., 2010; Tiefenbeck et al., 2018, 2019; Fang et al., 2020). We advance this strand of literature by examining the complementarities of goals and real-time feedback in a dynamic setting. By collecting real-time shower data, we can precisely estimate the effects of initial goal difficulty on conservation efforts, as well as the subsequent effects when goals are reset. Our unique setting further allows us to test the effectiveness of these behavioral interventions in the absence of monetary incentives and selection bias.

The rest of the paper is organized as follows. Section 2 describes our experimental design. Section 3 outlines the behavioral predictions of our treatments. Section 4 presents the data and provides descriptive analysis. Section 5 presents the main results. Section 6 concludes.

2 Experimental Design

We implemented a field experiment in two residential colleges, "Cinnamon" and "Tembusu", at NUS University Town, spanning from August 5, 2019 to November 24, 2019. This was in collaboration with the NUS Office of Housing Services, which had an interest in exploring behavioral interventions to promote resource conservation on campus.

2.1 Background

Each residential college comprises 21 storeys and houses over 500 bedrooms, accommodating local undergraduates, international exchange students, and a small group of faculty members. Our pool of subjects primarily comprised incoming freshmen, with faculty members excluded from the study. Figure A1 displays photos of the experimental site at Cinnamon and Tembusu colleges.

A total of 324 smart shower heads were installed in all designated bathrooms at Cinnamon and Tembusu colleges. Notably, each floor has two types of bathrooms: apartment and common bathrooms (see Figure 1). Residents who live in a shared apartment have access to their own apartment bathroom, while those who stay in single corridor rooms can only use the common

bathrooms.[6] From anecdotal evidence, residents typically store their toiletries in one particular bathroom, and hence it is safe to assume that the majority use the same bathroom for showers. In light of this, we chose to randomise at the residence × floor × bathroom type level. Each unit of randomization consists of between 4 and 6 shower heads that receive the same treatment assignment, shared by 18 residents on average.

The residents did not have to actively agree to participate in the study as the smart shower heads were installed in the bathrooms by NUS Office of Housing Services prior to them moving in. This rules out selection bias, whereby individuals with higher environmental awareness might be more likely to participate in studies on resource conservation. Again, we highlight that the residents have no monetary incentives to save water or energy as they pay a fixed monthly rent.

Figure 1: *Representative floor plan of Cinnamon and Tembusu colleges*

Notes. The figure shows the representative floor plan of both Cinnamon and Tembusu colleges. Every floor comprises two bathroom types, i.e. apartment bathrooms (in blue) and common bathrooms (in orange), each representing a unit of randomization at the residence × floor × bathroom type level. See https://uci.nus.edu.sg/ohs/future-residents/undergraduates/utown/room-types/ for further details.

[6] At the beginning of each academic year, students can opt for their preferred room type (i.e., single room in shared apartment or single corridor room) on either mixed or single-gender floor.

2.2 Real-Time Feedback Technology

The smart shower head is engineered by HYDRAO, a French water-technology and data startup. During a shower event, the smart shower head displays a colored light, which changes in real-time based on water usage. This provides users with immediate feedback on their shower water use. The exact thresholds and colored display can be configured remotely, which enables us to implement different goals for our treatment groups.

During each shower event, the smart shower head collects real-time data when connected to the server via WiFi. As a safeguard, it has an internal memory capable of storing up to 200 shower events. If real-time transmission fails, the data will be stored and transmitted as an offline shower event as soon as the WiFi connection is restored. If a shower is interrupted for up to 2 minutes (e.g., for soaping), the smart shower head considers it part of the same shower event. Beyond 2 minutes, it assumes that a new shower event has started.

In addition, the smart shower head is self-powered by water flow through a mini-turbine, eliminating the need for an external power supply. For home usage, a HYDRAO shower app is available to synchronize each shower head with a mobile phone for configuring the settings. For the purpose of our experiment, we remotely set the color thresholds for all shower heads using gateways to minimize disruptions for the residents. Importantly, our subjects were not informed of the app and unable to change the configured settings of the shower heads, ensuring the integrity of our randomization process.

2.3 Treatment Assignments

Our field experiment comprises three stages: baseline, phase 1 and phase 2. The baseline period was in effect for 6 weeks from the start of the semester (August 5, 2019 to September 15, 2019). Phase 1 of the intervention lasted for the subsequent 5 weeks, occurring from September 16, 2019 to October 21, 2019. We then proceeded to phase 2, which extended until the end of the semester, from October 22, 2019 to November 24, 2019.

During the baseline period, no interventions were implemented in the shower facilities. The main objective was to collect data on the residents' pre-experimental showering behavior. The baseline data include observable characteristics from each device (the term is used interchangeably with shower head), such as water use per shower, number of showers per day, and flow rate, which we used for randomization checks in the following section.

For the intervention, we assigned subjects to four experimental groups: a control group and

three treatment groups, and these assignments were fixed throughout the experiment. The Control group received neither the shower poster nor real-time feedback in both phases. The Moral Suasion (MS) group received a shower poster appealing to users to keep their water use below a specified level but received no feedback through the shower heads. The shower poster referenced a goal of 28L in phase 1, which changed to 24L in phase 2. The 18L goal group received a shower poster and real-time feedback corresponding to the goal of 18L in phase 1, and thereafter 24L in phase 2. Similarly, the 28L goal group received a shower poster and real-time feedback for the goal of 28L in phase 1, which was adjusted to 24L in phase 2. In summary, the treatment groups received different goals in phase 1 but the same goal in phase 2, with the information conveyed through the respective shower posters (see Figure A2).

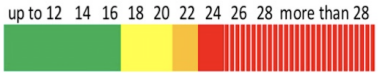
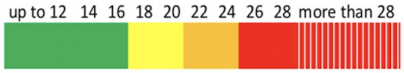
For the 18L goal and 28L goal groups, we programmed the shower heads to provide real-time feedback through colored lights (resembling the traffic light system) that correspond to a set of water use thresholds. At the start of each shower, the shower head displayed a green light, progressing to yellow, orange and red with increasing water usage. When the water volume exceeded the goal, the shower head would begin to display a blinking red light. Table 1 summarizes the key features of the experimental groups. For further illustration, Figure 2 depicts how a shower head, assigned to the 18L goal group, provides real-time feedback over the course of a shower event in phase 1.

Our experimental design allows us to identify the effect of moral suasion alone, and the marginal effect of real-time feedback (on top of moral suasion) under different goals. The assignment of different goals in phase 1 and a common intermediate goal in phase 2 further allows us to study how goal difficulty influences effort provision in a dynamic setting. Specifically, in phase 1, we compare the initial effects of a moderate goal (28L) versus a hard goal (18L) on shower water use. In phase 2, we examine whether initial goal assignments shape how individuals respond to the new intermediate goal (24L).

3 Behavioral Predictions

Our experimental setup allows us to test three predictions derived from the literature on goal-setting and reference-dependent preferences (Heath et al., 1999; Kahneman and Tversky, 1979; Tversky and Kahneman, 1991; Kőszegi and Rabin, 2006). If goals act as reference points, then the psychological utility of reaching the goal depends on loss aversion and diminishing

Table 1: *Summary of treatment assignments*

Group \ Stage	PHASE 1	PHASE 2
Control	none	none
Moral Suasion	poster only (referencing 28L goal)	poster only (referencing 24L goal)
18L GOAL	poster + feedback (referencing 18L goal) 	poster + feedback (referencing 24L goal) 
28L GOAL	poster + feedback (referencing 28L goal) 	poster + feedback (referencing 24L goal) 

Notes. All the experimental groups received neither the shower poster nor real-time feedback in the baseline period. For the 18L goal and 28L goal groups, the shower poster is augmented with information explaining how the shower head changes colors with the corresponding thresholds.

Figure 2: *Implementation of 18L goal group*

(a) *0L to 12L* **(b)** *12L to 14L* **(c)** *14L to 16L* **(d)** *16L to 18L*

Notes. In panel (a), the shower head is displaying a green light, indicating that water use up to that particular point is below 12 liters. The remaining panels are self-explanatory. Beyond 18 liters of water use (i.e., the assigned goal in phase 1), the shower head starts to display a blinking red light.

sensitivity around it. First, goals could introduce a level difference between treatments in phase 1: an ambitious goal may elicit strong effort, because it puts an individual firmly in the loss domain and mitigating the loss has a high marginal value. However, there is a counteracting force: if goals are too hard, they may elicit lower effort from individuals because of diminishing sensitivity relative to the reference point.

In phase 1 of our experiment, we chose the 28L and 18L goals to represent the moderate and hard goals, respectively: the 28L goal was set such that it was attainable with reasonable effort, as suggested by treatment effects of real-time feedback found in other studies (Agarwal et al., 2018; Fang et al., 2020). By contrast, the 18L goal was set to be extremely difficult to meet given baseline shower behavior by that standard.[7] We thus arrive at the following prediction:

Hypothesis 1: ***Initial goal difficulty and effort.*** *In phase 1, there is a level difference in treatment effects between the two treatment groups: conservation efforts in the 28L goal group are greater than in the 18L goal group.*

We set the goals such that the above hypothesis would hold on average, but any level difference in treatment effects is expected to be less pronounced among the low baseline users. For these users, the 18L goal becomes more realistic and thus may elicit greater effort. In contrast, the 28L goal may be close to their typical usage, placing these users in the gain domain where the marginal value of effort is lower, potentially leading to reduced effort.

At least two different mechanisms of how goals affect reference points could lead to the prediction of Hypothesis 1. It may be that goals directly serve as reference points and thus create the above pattern, as is argued in Heath et al. (1999). However, it could also be that reference points are driven by recent expectations, as in Kőszegi and Rabin (2006), and goals merely serve to influence those expectations.

However, this is no longer the case for the predictions regarding phase 2. The main feature we are interested in testing is how individuals respond when the difficulty of goals is changed over time. In phase 2 of our experiment, for the 28L goal group, we introduce a tougher goal of 24L. In contrast, for the 18L goal group, the same 24L goal constitutes an easier goal. How the two groups respond to the new goal depends on whether and how goals affect reference points. If goals directly act as reference outcomes, then the introduction of a common 24L goal in phase 2 should lead to the same reference point, and hence to the same outcomes for both groups. This is summarized in our next prediction:

[7] The average shower water use in the baseline period is 31.9 liters, with standard deviation of 23.0.

Hypothesis 2a: ***Goals as direct reference points.*** *If goals act directly as reference points, then outcomes in phase 2 should be identical for the 28L goal and 18L goal groups.*

Notwithstanding, if recent expectations or lagged outcomes shape individuals' reference points, changing goals over time could have different effects. For the 28L goal group, the initial moderate goal encourages conservation by creating an immediate loss (relative to baseline consumption), but is not too difficult to reach. As the goal can be reached, the new outcome sinks in as the reference point (or the expectation of the outcome), and again pushing individuals into the loss domain in phase 2. This raises the marginal benefit of conservation, thus increasing conservation efforts even further. By contrast, moving from the 18L goal to the 24L goal makes the goal easier relative to the previous benchmark. This shift potentially moves individuals into the gain domain in phase 2, thus reducing their marginal benefit from conserving water and leading to lower conservation efforts. We summarize this as follows:

Hypothesis 2b: ***Lagged expectations of outcomes affect reference points.*** *If reference points are affected by lagged outcomes or recent expectations, the conservation efforts between the 18L goal and 28L goal groups will diverge in phase 2: treatment effects of the 28L group will directionally strengthen, while treatment effects of the 18L group will directionally weaken.*

This divergence in treatment effects should again be related to the goal difficulty effects from Hypothesis 1 and more pronounced for high-baseline users: high-baseline users in the 18L treatment may have come to expect to fall short of the goal by a lot. Easing the goal to 24L may thus reduce their motivation to conserve. By contrast, a comparable individual assigned to the 28L condition well meet the goal in phase 1. Tightening the goal to 24L now creates a sensation of a loss, thus increasing the motivation to conserve.

4 Data and Descriptive Analysis

Our data were collected from the field experiment described above, resulting in a total of 128,323 recorded shower events from 301 smart shower heads over a 16-week period.[8] The shower events can be categorized as live showers and offline showers. Live showers were those where data were transmitted in real-time from the shower heads to our server, providing precise information about the date and time of each shower event. On the other hand, offline showers had a time

[8] We excluded observations that recorded water use of 4 litres or less as these instances are unlikely to be actual shower events, but rather for other purposes like cleaning.

lag in data transmission, making it challenging to pinpoint the exact occurrence of each shower event. To increase precision of our estimates, we thus consider the sample of 115,992 live shower events (comprising 90% of all recorded showers) from 284 shower heads as our primary data for analysis.[9]

In the baseline period, we observe a stable pattern of around 1,300 live showers on a regular weekday, and about half the number on weekends. On the intensive margin, there is a right-skewed distribution of shower water use, with a median of 27.0L and a mean of 31.9L (see Figure A3).

4.1 Randomization Checks

To begin, we perform balance tests to support the integrity of the randomization. Table 2 presents a comparison of the treatment groups to the control group based on key observable characteristics during the baseline period. It is apparent that balance of treatment is attained as almost all observables, especially water use per shower and total number of showers, show no significant differences across groups. There is only slight statistical difference in the number of days since last transmission between the control and 28L goal group.[10] However, this discrepancy can be attributed to a single shower head in the control group that rarely records live shower events; thus, it does not constitute a cause for concern.[11] We conclude that our experimental groups are well-balanced, and any observed differences during the intervention can be interpreted as causal treatment effects.

4.2 Descriptive Evidence

We compare how our interventions influence shower behavior under different goals, considering both the intensive margin (i.e., water use per shower) and the extensive margin (i.e., daily number of showers). We divide the full sample into three time periods: baseline, phase 1 and phase 2. Recall that in phase 1, the 18L goal and 28L goal groups received real-time feedback referencing different goals (18L vs. 28L), and later in phase 2, both groups converged to the common goal of 24L.

[9] Our main results are qualitatively and quantitatively similar using the full sample of recorded shower events, though with slightly less precision due to the need for interpolation of the dates and times for the offline shower events. The results are available upon request.

[10] The variable `days since last transmission` is defined as the number of days since a shower head last transmitted shower data to our server at the end of the baseline period.

[11] This shower head last transmitted shower data 24 days before the start of phase 1, possibly due to poor Wi-Fi coverage in the bathroom. Out of the 238 observations recorded by the shower head during the entire

Table 2: *Randomization checks*

Dependent variable:	Baseline averages by shower head						
	Water use per shower (in litres) (1)	Total number of showers (2)	Duration per shower (in seconds) (3)	Fraction of *live* showers (4)	Days since last transmission (5)	Suite bathroom (6)	Floor (7)
Moral Suasion	−1.595	12.819	−21.511	−0.001	−0.183	−0.074	0.573
	(1.797)	(13.452)	(17.098)	(0.025)	(0.591)	(0.173)	(1.785)
18L goal	−0.845	17.763	5.060	0.028	−0.323	0.087	1.259
	(2.095)	(14.899)	(22.689)	(0.020)	(0.386)	(0.166)	(1.650)
28L goal	−2.300	17.412	−18.388	0.013	−0.687**	−0.008	−0.354
	(1.585)	(12.879)	(17.585)	(0.023)	(0.371)	(0.174)	(2.005)
Constant	33.546***	140.955***	375.328***	0.887***	0.746**	0.567***	12.075***
	(1.301)	(8.123)	(14.063)	(0.017)	(0.370)	(0.123)	(1.319)
p-value for F-test	0.522	0.476	0.413	0.354	0.004	0.811	0.791
R^2	0.009	0.009	0.018	0.011	0.015	0.014	0.015
Observations	284	284	284	284	284	284	284

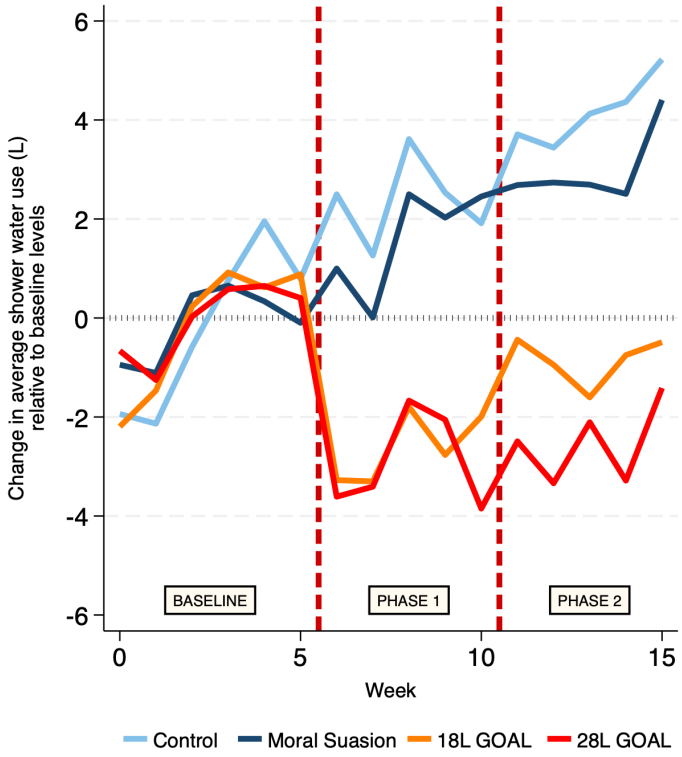
Notes. The results are obtained by regressing the various baseline averages of observables on assigned experimental groups. The omitted group is the control group. Standard errors clustered at the residence × floor × bathroom type level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3 depicts how our treatments impacted average shower water use relative to baseline levels. During the baseline period, we observe that all experimental groups exhibit similar trends in mean shower water use, consistent with our randomization checks. However, notable differences emerge during the intervention phases. The control group shows a visible upward trend in mean shower water use over time, while the MS group shows a slightly better, but otherwise similar, trend in both phases. In contrast, the 18L goal and 28L goal groups experience sharp reductions in mean water use per shower with the onset of real-time feedback, coupled with moral suasion, in phase 1. These substantial effects persist in phase 2, though their magnitudes vary based on the initial assigned goal (18L vs. 28L). We delve into the dynamics of these treatment effects in greater detail below.

Analogously, Figure 4 depicts how our treatments impacted average daily number of showers relative to baseline levels. On the extensive margin, we observe a similar trend across all experimental groups, with no significant change in average levels during both phases of the intervention. These results indicate that our interventions have little to no effect on the extensive margin, a finding that will be formally verified in the next section. The only anomaly is a noticeable drop in the daily number of showers in week 7, which coincides with the recess week when some residents are away from campus for a one-week break. It is reassuring that we observe the

experiment, only 11% are classified as live showers.

Figure 3: *Intensive margin of shower behavior by experimental groups*

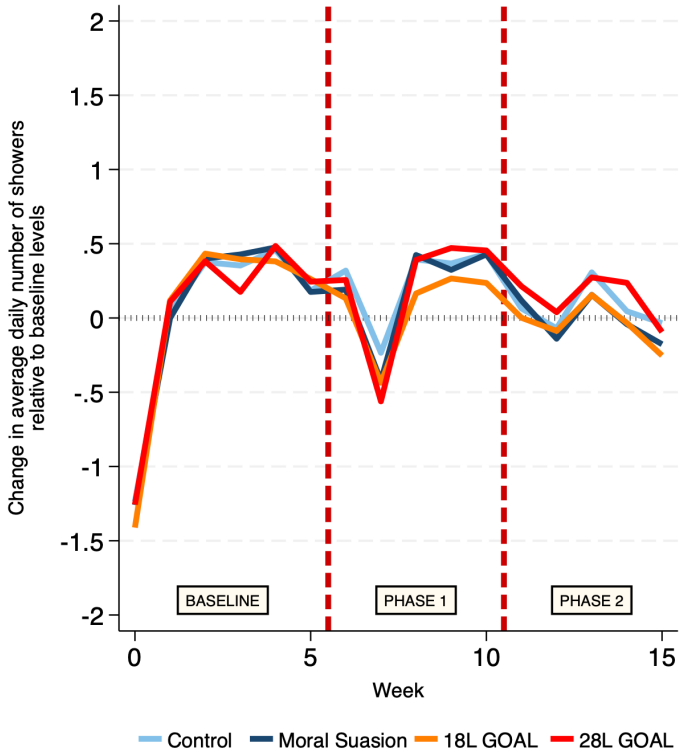


Notes. Shower water use (in litres) is averaged across all shower heads within the same experimental group on a weekly level, with the mean baseline levels subtracted. To reduce the influence of outliers, we exclude observations that recorded under 4 litres (inclusive) of water use and top-code values exceeding 200 litres, as these instances are unlikely to be actual shower events.

same dip in levels across all four experimental groups, suggesting no differential selection out of the experiment during the mid-term break.

Next, Figure 5 displays the difference-in-differences estimates of the ATEs for the MS, 18L goal, and 28L goal groups, examining the impact of our treatments on the intensive and extensive margins of shower behavior, respectively. In panel (a), we observe a modest decrease in mean water use for the MS group in phase 1. In contrast, the 18L goal and 28L goal groups, which received real-time feedback in addition to moral suasion, exhibit sharp reductions in water use. The standard error bars around the means suggest highly significant effects. Interestingly, both groups appear to respond similarly to the treatment, despite having different initial goals. However, when both groups converged to the common 24L goal in phase 2, we observe a divergence in treatment effects. Once again, the use of moral suasion alone shows only marginal effects, if any, on shower water use. Turning to panel (b), which displays the ATEs on the extensive margin, we observe no significant effect on the daily number of showers in both phases of the

Figure 4: *Extensive margin of shower behavior by experimental groups*

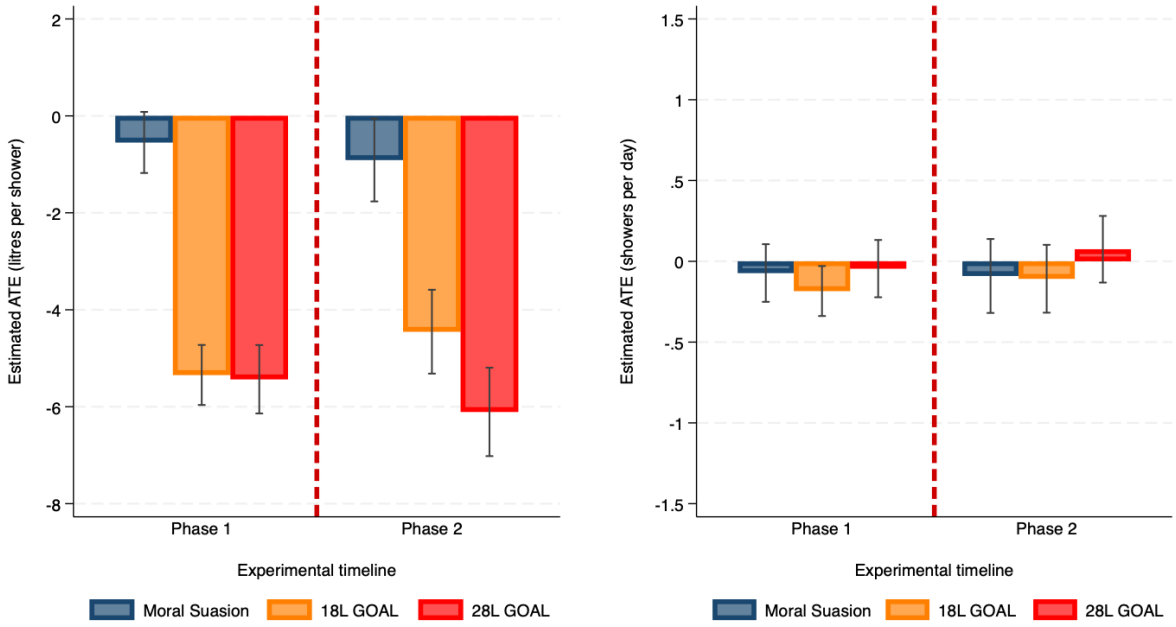


Notes. Daily number of showers is averaged across all shower heads within the same experimental group on a weekly level, with the mean baseline levels subtracted. To reduce the influence of outliers, we exclude observations which recorded under 4 litres (inclusive) of water use, as these instances are unlikely to be actual shower events.

intervention. This finding suggests that there is no differential selection into or out of the experiment among our treatment groups. By ruling out selection effects (changes in the composition of subjects), we can attribute any changes on the intensive margin as the sole response to our respective treatments.

To enrich our analysis, we split the sample (of shower heads) by whether the average water use per shower during the baseline period was above or below the median of 30.7L. Figure 6 presents the estimated ATEs of water use per shower for the experimental groups: panel (a) displays the results for above-median users, while panel (b) displays the results for below-median users. In panel (a), we observe that the dynamic effects of goal-setting are particularly pronounced for above-median users. In phase 1, both the 18L goal and 28L goal groups significantly reduced average water use to a similar extent. However, in phase 2, there is a clear divergence of ATEs between the two groups, driven by a significant weakening of the treatment effect in the 18L goal group. In contrast, we do not observe the same dynamics for below-median users. In both

Figure 5: *Estimated ATEs by experimental groups*



(a) *Intensive margin*

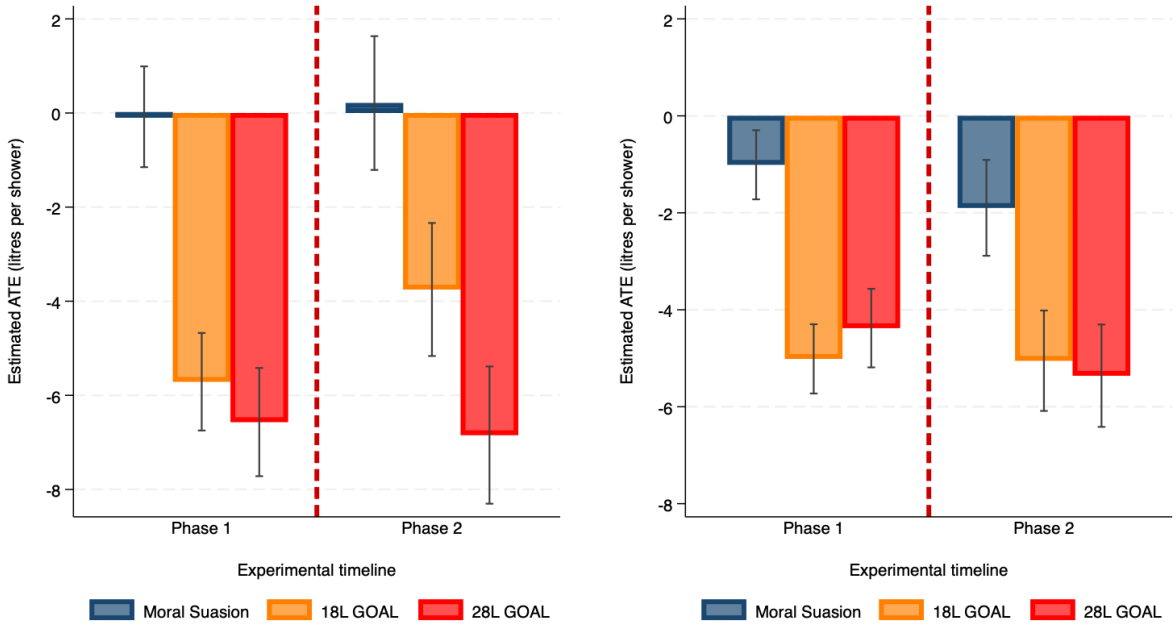
(b) *Extensive margin*

Notes. Each bar represents the difference-in-differences estimates of the outcome of interest for the experimental groups in phase 1 and 2, respectively, relative to the control group. Panel (a) looks at the intensive margin with water use per shower as the outcome variable, while panel (b) looks at the extensive margin with number of showers per day as the outcome variable. The error whiskers display +/− one standard error of the mean. To reduce the influence of outliers, we exclude observations that recorded under 4 litres (inclusive) of water use and top-code values exceeding 200 litres, as these instances are unlikely to be actual shower events.

phases, the reductions of average water use for the 18L goal and 28L goal groups are similar in magnitude.

For completeness, we also present the estimated ATEs of the daily number of showers for above-median and below-median users, respectively. No significant differences between the treatment groups were observed in both phases of the intervention, and the result is included in the Appendix (see Figure A4).

Figure 6: *Estimated ATEs (intensive margin) for above and below median users*



(a) *Above-median users*

(b) *Below-median users*

Notes. Each bar represents the difference-in-differences estimates of water use per shower for the experimental groups in phase 1 and 2, respectively, relative to the control group. Panel (a) displays the estimated ATEs for the sample of above-median users, while panel (b) displays the estimated ATEs for the sample of below-median users. The error whiskers display +/− one standard error of the mean. To reduce the influence of outliers, we exclude observations that recorded under 4 litres (inclusive) of water use and top-code values exceeding 200 litres, as these instances are unlikely to be actual shower events.

5 Main Results

This section details our empirical strategy and presents the main results of our experiment.

5.1 Effects of Moral Suasion and Real-Time Feedback Under Different Goals

To identify the causal effects of our treatments, we estimate the following linear model with three levels of fixed effects:

$$Y_{ith} = \alpha_i + \lambda_t + \gamma_h + \left(\beta_{MS1}\text{MS}_i + \beta_{18L1}\text{18L}_i + \beta_{28L1}\text{28L}_i\right) \times \text{PHASE1}_{ith} + \left(\beta_{MS2}\text{MS}_i + \beta_{18L2}\text{18L}_i + \beta_{28L2}\text{28L}_i\right) \times \text{PHASE2}_{ith} + \epsilon_{ith} \quad (1)$$

Y_{ith} represents the relevant outcome variable (such as water use per shower) for device i on day t and hour h. α_i is the device fixed effect, λ_t is the day fixed effect and γ_h is the hour-of-day fixed effect. MS_i is an indicator variable that equals one for the MS, 18L goal and 28L goal groups that all receive moral suasion, in the form of a shower poster. The 18L_i and 28L_i are indicator variables for being assigned to the 18L goal and 28L goal groups, respectively.[12] Similarly, PHASE1_{ith} is an indicator variable that equals one for the period when the initial goals (i.e. either 18L or 28L) were introduced, whereas PHASE2_{ith} is an indicator variable that equals one for the latter period when the shower goal is changed to 24L.[13] ϵ_{ith} is the random error term and standard errors are clustered at the residence $\times$ floor $\times$ bathroom type level, which is the unit of randomization.

Motivated by previous studies that have documented larger conservation effects among high baseline users (Ferraro and Price, 2013; Allcott, 2011; Tiefenbeck et al., 2018), we enrich our analysis by estimating the same specification (1) separately on the sample of above-median and below-median users. Our preferred specification includes device fixed effects to account for time-invariant differences across residents who are assigned to different experimental groups, as well as day and hour-of-day fixed effects to control for aggregate patterns in weather and lifestyle over the course of the experiment. The coefficients of interest are the respective β terms, which represent difference-in-differences estimates for each of the treatments relative to the control. These ATEs are identified from within-device variation over time, while controlling for aggregate hourly and daily shocks. To clarify, β_{MS1} and β_{MS2} represent the ATE of moral suasion in phase 1 and phase 2, respectively. Similarly, the coefficients for the interacted 18L_i and 28L_i variables

[12] Note that the MS_i variable is defined as one for all three treatment groups, not just the MS group. The reason for this is that both the 18L goal and 28L goal groups also receive moral suasion, and therefore the MS group serves as the relevant comparison for identifying the marginal effects of real-time feedback (in addition to moral suasion).

[13] To be precise, for the MS group, PHASE1_{ith} equals one from 4PM, September 16 to 5PM, October 22, whereas PHASE2_{ith} equals one from 5PM, October 22 onwards. For the 18L goal and 28L goal groups, we use the exact date and time the feedback was in place to define the PHASE1_{ith} and PHASE2_{ith} variables, respectively.

identify the additional effect of real-time feedback beyond moral suasion in each phase.

5.1.1 Effect on the intensive margin

We begin by examining how subjects adjusted their showering behavior on the intensive margin, using water use per shower as the outcome variable. Table 3 presents the results, with column 1 displaying the results from the full sample, and columns 2 and 3 showing analogous results from the samples of above-median and below-median baseline users, respectively.

Result 1: *In phase 1, conservation efforts in the 18L goal group and 28L goal group are not significantly different.*

In column 1, we find that moral suasion alone did not lead to a significant effect on water use per shower in either phase, although the point estimates of −0.6L and −0.8L are in the desired direction. On the other hand, the provision of real-time feedback, beyond the effect of moral suasion, induced substantial conservation effects of approximately 15% in phase 1 (i.e., reduction of between 4.8 and 4.9 litres per shower). These findings are consistent with previous studies on water conservation involving smart shower heads (Tiefenbeck et al., 2018; Goette et al., 2019). Interestingly, we did not detect any significant difference in the ATEs of real-time feedback for different assigned goals (18L vs. 28L) in phase 1 ($p = 0.919$). This result runs counter to Hypothesis 1, which posits that the 28L goal group would exhibit greater conservation efforts than the 18L goal group, since the moderate goal was expected to be attainable with reasonable effort.

Result 2: *In phase 2, conservation efforts in the 18L goal group and 28L goal group pull apart in opposite directions.*

With the goal adjustment to the common level of 24L in phase 2, a divergence of the ATEs becomes apparent. The change in treatment effects from phase 1 to phase 2 is significantly different between the 18L and 28L conditions ($p = 0.006$), and this is mainly driven by underperformance of the 18L goal group. As evident in column 1, the conservation efforts of the 18L condition substantially declined in phase 2 compared to phase 1, with the ATE weakening from −4.8L to −3.7L ($p = 0.059$). Conversely, the conservation efforts of the 28L condition strengthens from −4.9L to −5.4L with the tightening of the goal, although this difference is not statistically significant ($p = 0.455$). In fact, the pulling apart of the conservation efforts in both

groups is substantial enough that the ATEs in phase 2 are now significantly different ($p = 0.070$). These documented effects are quantitatively significant, with the difference in treatment effects in phase 2 amounting to 1.7L per shower, which is approximately 30% of the ATE of the 28L goal group during phase 2. Taken together, our data reject Hypothesis 2a, which posits that the shower goals directly act as reference points, as this interpretation would have implied the same outcomes for both groups in phase 2. In contrast, we find strong support for Hypothesis 2b, which suggests that reference points are influenced by recent expectations or lagged outcomes, and hence we would observe the conservation efforts in the 18L goal and 28L goal groups pulling apart in the predicted directions.

Result 3: *There is substantial treatment heterogeneity by baseline water use, with the effects of changing goals being particularly pronounced among above-median baseline users.*

When the full sample is split by above- and below-median baseline water use, we still observe the same directional effects of goal adjustments in phase 2: the 18L goal group exhibits reduced performance while the 28L goal group intensifies conservation efforts with the new 24L goal. Notably, the divergence in conservation efforts during phase 2 is especially pronounced among those with above-median baseline use. As presented in column 2, the ATEs for both the 18L and 28L conditions in phase 1 are similar ($p = 0.349$), but there is a striking difference in the ATEs in phase 2 ($p = 0.014$). This is primarily driven by a significant weakening of the ATE for the 18L condition, declining from -5.5L to -4.0L, as the initial 18L goal is relaxed to 24L in phase 2. In contrast, the effects are more muted among those with below-median baseline use, and do not attain statistical significance.

5.1.2 Effect on the extensive margin

Next, we examine how subjects adjusted their showering behavior on the extensive margin, using number of showers per day as the outcome variable. Considering that real-time feedback, coupled with moral suasion, induced substantial conservation effects of approximately 14% – 19%[14], it is pertinent to assess whether these savings were offset by subjects taking more showers each day. Conversely, if the treatments had prompted subjects to reduce their daily shower frequency, it could potentially lead to negative externalities in the form of hygiene problems. In addition,

[14]This corresponds to between 4.5 and 6.2 litres per shower, as evident from Table 3. The lower bound of 4.5 litres per shower is given by the sum of estimates $\hat{\beta}_{MS2}$ and $\hat{\beta}_{18L2}$, while the upper bound of 6.2 litres is given by the sum of estimates $\hat{\beta}_{MS2}$ and $\hat{\beta}_{28L2}$.

Table 3: *Effects of moral suasion and real-time feedback on water use per shower*

Dependent variable:	Water use per shower (liters)					
	Full Sample (1)		Above Median (2)		Below Median (3)	
	PHASE 1	PHASE 2	PHASE 1	PHASE 2	PHASE 1	PHASE 2
MS × PHASE	−0.570	−0.829	−0.240	0.308	−0.832	−1.707**
	(0.528)	(0.869)	(0.601)	(1.227)	(0.666)	(0.845)
18L × PHASE	−4.836***	−3.703***	−5.538***	−3.953***	−4.154***	−3.431***
	(0.584)	(0.870)	(0.745)	(1.131)	(0.612)	(0.864)
28L × PHASE	−4.906***	−5.386***	−6.473***	−7.179***	−3.444***	−3.714***
	(0.625)	(0.844)	(0.944)	(1.189)	(0.553)	(0.819)
Constant	33.845***		39.693***		28.211***	
	(0.271)		(0.354)		(0.285)	
p-values of interest:						
Level differences in TEs: $\beta_{18L} = \beta_{28L}$	0.919	0.070	0.348	0.014	0.275	0.729
Divergence in TEs: $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$	0.006		0.006		0.097	
Change of TE over time: $\beta_{18L1} = \beta_{18L2}$	0.059		0.042		0.349	
Change of TE over time: $\beta_{28L1} = \beta_{28L2}$	0.455		0.436		0.717	
F-test	0.000		0.000		0.000	
Device FEs	✓		✓		✓	
Date FEs	✓		✓		✓	
Hour-of-day FEs	✓		✓		✓	
R^2	0.139		0.093		0.073	
Observations	115,992		56,574		59,418	

Notes. This table shows the effects of moral suasion and real-time feedback on water use per shower under different goals in each phase. The results are obtained by estimating equation (1) that includes controls for device, day and hour dummies. For ease of interpretation, we split the estimates from a single regression into 2 sub-columns, by each phase of the intervention. To reduce the influence of outliers, we exclude observations that recorded under 4 litres (inclusive) of water use and top-code values exceeding 200 litres, as these instances are unlikely to be actual shower events. Standard errors clustered at the residence × floor × bathroom type level in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

there may be attrition bias, where subjects drop out of the study non-randomly. To formally test this, we re-estimate equation (1) without hour-of-day fixed effects, this time using number of showers per device per day as the outcome variable.

Table 4 presents the results, with column 1 displaying estimates from the full sample, and columns 2 and 3 showing the corresponding estimates from the samples of above-median and below-median baseline users, respectively. We observe that almost all point estimates are statistically insignificant, which aligns with the descriptive evidence presented earlier (see Figure 4). Across all three columns, we firmly fail to reject the null hypothesis of equal treatment effects between the 18L and 28L conditions in either phase. This allays our primary concern about subjects compensating for reduced water use per shower by taking more showers each day. Consequently, we conclude that our treatments only induced adjustments on the intensive margin, allowing us to focus our discussion on this aspect.

5.2 Non-Parametric Estimation of Heterogeneous Treatment Effects

Our regression analysis above suggests that the treatment effects of changing goals (at the intensive margin) are moderated by baseline water usage. In particular, we documented that setting too hard an initial goal can have the unintended effect of diminishing conservation efforts when the goal was subsequently relaxed, and these effects were especially pronounced among high baseline users.

To gain deeper insights into the underlying treatment heterogeneity, we now apply the cluster-robust causal forest algorithm to our dataset, following Athey and Wager (2019). This is a non-parametric, machine learning-based method of Athey et al. (2019) which extends the random forest algorithm of Breiman (2001) to heterogeneous treatment effect estimation. Under modest regularity conditions, the causal forest algorithm produces estimates that are consistent and asymptotically normal (Chernozhukov et al., 2018). In addition, the algorithm allows us to systematically examine which are the covariates that moderate the treatment effects, without having to pre-specify a particular covariate of interest as in a parametric model. Before proceeding with our application, we briefly explain how the causal forest works.

We begin by introducing some new notation for ease of exposition. We denote the outcome of interest by Y, the treatment assignment by W and the vector of covariates by X. Following the potential outcomes framework, we posit the existence of potential outcomes, $Y_i(0)$ and $Y_i(1)$, and assume that we observe $Y_i = Y_i(W_i)$ in our data. The conditional average treatment effect

Table 4: *Effects of moral suasion and real-time feedback on number of showers per day*

Dependent variable:	Number of showers per day					
	Full Sample (1)		Above Median (2)		Below Median (3)	
	PHASE 1	PHASE 2	PHASE 1	PHASE 2	PHASE 1	PHASE 2
MS × PHASE	−0.261*	−0.138	−0.410	−0.350	−0.107	0.065
	(0.136)	(0.133)	(0.249)	(0.215)	(0.135)	(0.222)
18L × PHASE	−0.007	0.061	0.204	0.214	−0.232*	−0.106
	(0.109)	(0.108)	(0.184)	(0.204)	(0.126)	(0.212)
28L × PHASE	0.055	0.108	0.306	0.262	−0.203	−0.061
	(0.145)	(0.158)	(0.272)	(0.276)	(0.144)	(0.215)
Constant	4.250***		4.169***		4.339***	
	(0.055)		(0.093)		(0.055)	
p-values of interest:						
Level differences in TEs: $\beta_{18L} = \beta_{28L}$	0.658	0.760	0.686	0.848	0.799	0.796
Divergence in TEs: $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$	0.898		0.742		0.908	
F test: $\beta_{MS} = \beta_{18L} = \beta_{28L} = 0$	0.190	0.750	0.379	0.443	0.007	0.968
Overall F-test	0.144		0.777		0.002	
Device FEs	✓		✓		✓	
Date FEs	✓		✓		✓	
R^2	0.550		0.536		0.567	
Observations	27,817		13,924		13,893	

Notes. This table shows the effects of moral suasion and real-time feedback on number of live showers per day under different goals in each phase. The results are obtained by estimating a variant of equation (1) that includes controls for device and day dummies. For ease of interpretation, we split the estimates from a single regression into 2 sub-columns, by each phase of the intervention. To reduce the influence of outliers, we exclude observations which recorded under 4 liters (inclusive) of water use, as these instances are unlikely to be actual shower events. Standard errors clustered at the residence × floor × bathroom type level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(CATE) function is defined as $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$, and is the estimand of interest for heterogeneous treatment effects. We also define the auxiliary quantities, namely the propensity score $e(x) \equiv \mathbb{E}[W_i|X_i = x]$ and the conditional mean function $m(x) = \mathbb{E}[Y_i|X_i = x]$.

Recall in the regression case, each decision tree is grown by recursively partitioning the covariate space into two regions at each split, such that the weighted difference in outcome of the subgroups is maximized. Thereafter, we obtain the random forest by growing and averaging the predicted outcomes of many such regression trees. The idea of the causal forest is closely related to the random forest; for each tree, we instead recursively partition the covariate space with a splitting rule that maximizes the weighted difference in treatment effect of the two subgroups. However, unlike the observed outcomes, we do not observe the true treatment effect τ in our data, and therefore need to estimate it. The approach taken by Athey et al. (2019) is to assume constant treatment effects within each region for each candidate split. This amounts to solving a partially linear model, for which we can apply Robinson's transformation and estimate τ with a residual-on-residual regression (Robinson, 1988). Algorithmically, we first obtain an estimate of the propensity score $\hat{e}(\cdot)$ by predicting W_i from X_i and the conditional mean function $\hat{m}(\cdot)$ by predicting Y_i from X_i, using any modern supervised machine-learning method (e.g. random forest, boosted tree, neural net, lasso, etc.). Then, we obtain the estimated treatment effect $\hat{\tau}$ by regressing the residual $Y_i - \hat{m}(X_i)$ on the residual $W_i - \hat{e}(X_i)$. Finally, we need a way of aggregating the causal estimates of $\hat{\tau}(x)$ across trees that is robust to regions that may be highly variable. We refer the interested reader to Athey et al. (2019) for more details.

In our application of causal forests, we expect the observations to be clustered, in that the outcomes of interest (i.e. water use per shower) are correlated within the same bathroom. Therefore, to estimate the CATE function, we proceed as follows: For each treatment group $g \in \{MS, 18L, 28L\}$ and each phase $t \in \{1, 2\}$, we construct a subsample $\{(Y_i, W_i, X_i)\}_{g,t}$ consisting of all observations belonging to the control and treatment group g, from the baseline period and phase t. $Y_i \in \mathbb{R}$ is the outcome of interest, i.e. water use per shower. W_i is a treatment dummy that equals one for observations from treatment group g, and zero for the control. $X_i \in \mathbb{R}^p$ with $p = 31$ is a vector of covariates comprising mean baseline water use, floor, bathroom type, hour of day, etc. In particular, X_i includes a dummy variable '*Post*' that equals one for observations from phase $t \in \{1, 2\}$, and zero for the baseline period.

For each subsample $\{(Y_i, W_i, X_i)\}_{g,t}$, we estimate the CATE function of treatment group g in phase t by running the cluster-robust causal forest analysis (see Algorithm 1 in Athey and

Wager (2019) for exact implementation). As a robustness check, we also estimate the CATE function for each treatment group in the baseline period, relative to the control. Since there was no intervention in the baseline period, we would expect the estimated CATE function to be essentially zero for every treatment group; we would henceforth term it as the pseudo-CATE function.[15]

Figure 7 presents the results, categorized by treatment group and phase.[16] It is worth noting that the pseudo-CATE function for each treatment group in the baseline period is depicted by the light blue points within each subfigure. Across all six subfigures, the cluster of light blue points is centred around zero, with slightly negative estimates observed in the regions of high baseline water use. This validates the use of causal forests on our dataset to estimate heterogeneous treatment effects.

We first interpret the subfigures corresponding to phase 1. For the MS group and 18L goal group, the estimated CATE functions appear to be constant, with little indication of heterogeneity by mean baseline water use (Figures 8a, 8c). For the 28L goal group, the estimated CATE function provides strong evidence of heterogeneity, with the treatment effects increasing linearly with mean baseline water use (Figure 8e). The results indicate that, in the 28L condition, individuals with high baseline consumption exerted more effort to reduce water use in phase 1, relative to their counterparts in the 18L condition. This suggests that setting too hard an initial goal has a demotivating effect for the high baseline users, and diminishes their conservation efforts from the outset.

Next, we turn to the subfigures corresponding to phase 2. For the MS group, the overall pattern appears similar to that observed in phase 1 and if anything, the treatment effects appear to slightly weaken with mean baseline water use (Figure 8b). For the 18L goal group, there appears to exhibit some treatment heterogeneity, albeit in a markedly non-linear manner (Figure 8d). For the 28L goal group, we continue to observe strong evidence of treatment heterogeneity, though the estimated CATE function no longer appears to be linear (Figure 8f).

The cluster-robust causal forest analysis allows us to directly examine how the CATE function changes when the goal is reset to 24L in phase 2. We see a distinct drop in the estimated CATEs (i.e., stronger treatment effects) around the neighborhood of 30L for both the 18L goal

[15] To the extent that the control group has slightly higher mean baseline water use, albeit not significantly different from the treatment groups (see Table 2), we may also expect the estimated pseudo-CATE to be slightly negative.

[16] We do not find meaningful treatment effect heterogeneity with respect to other covariates, and hence, we focus on the effect modification along mean baseline water use.

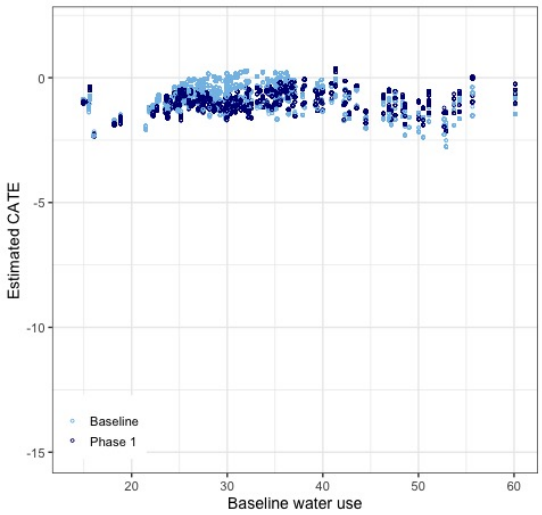
and 28L goal groups. In particular, we observe two counteracting forces in the 18L goal group: individuals with baseline consumption below 30L exert a higher conservation effort relative to phase 1, but those with baseline consumption above 35L reduce their conservation efforts even more. By contrast, tightening the goal for the 28L goal group increases the conservation efforts of individuals with baseline consumption around 30L, and leads only to a slight drop-off for very high baseline individuals. This set of results supports our interpretation that setting too hard a goal can permanently diminish motivation and effort provision, even after adjusting the initial goal to an intermediate level.

6 Conclusion

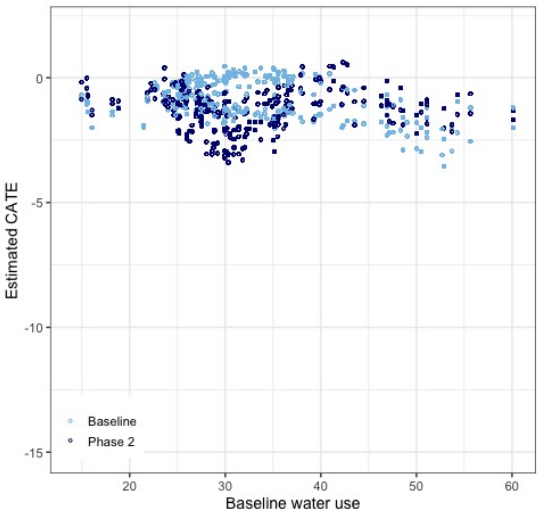
In this paper, we implemented a natural field experiment to examine the dynamic effects of goal-setting on effort provision in the context of resource conservation. Our two key treatment groups started with different goals, but midway through the intervention, had their respective goals adjusted to the same level of difficulty. In a setting with no pecuniary incentives and self-selection bias, we find that the provision of real-time feedback (in addition to moral suasion) induced large and significant reductions in shower water use. Notably, in phase 1 of our intervention, both the 18L goal and 28L goal groups performed equally well on average, despite being assigned different goals. However, in phase 2, when the initial goals are adjusted to the same level of difficulty, a divergence in their performances emerges: the 18L goal group now demonstrates diminished conservation efforts in comparison to the 28L goal group. This difference in phase 2 is quantitatively significant and amounts to an approximate 30% reduction in conservation efforts we observe in the 28L goal group. Moreover, the fact that the ATEs are similar in phase 1 masks substantial heterogeneous treatment effects. Employing non-parametric techniques, we find differential effect modification along baseline water use between the 18L goal and 28L goal groups throughout the intervention. In phase 1, high baseline users exert stronger conservation efforts under the moderate goal compared to the hard goal. Strikingly, this heterogeneous treatment effect carries over to phase 2, even though both groups are now assigned the same goal. This suggests that setting too hard an initial goal can permanently diminish effort provision and performance, which is a novel finding that has not been documented in the literature and goes beyond our behavioral predictions.

Several mechanisms can generate the observed behavior in our data. It could either be due

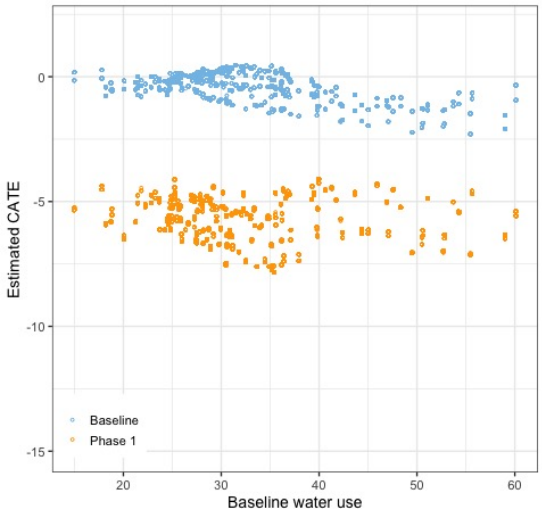
Figure 7: *Estimated CATE functions*



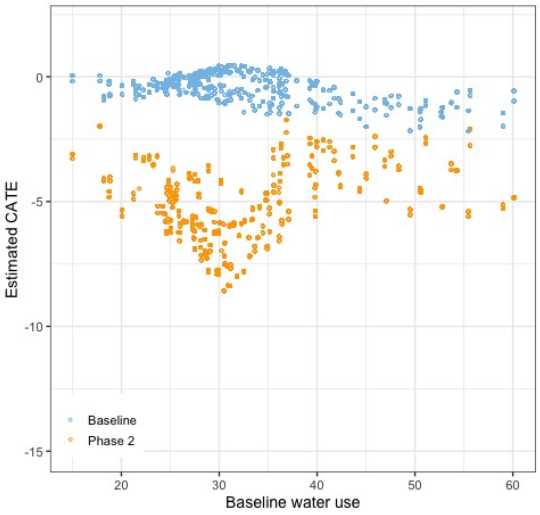
(a) *Moral Suasion group in Phase 1*



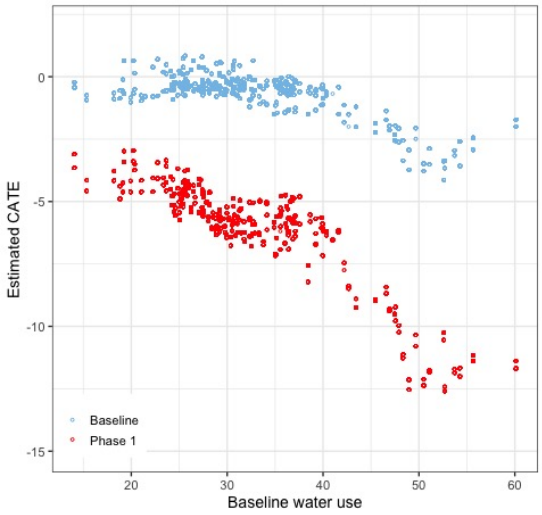
(b) *Moral Suasion group in Phase 2*



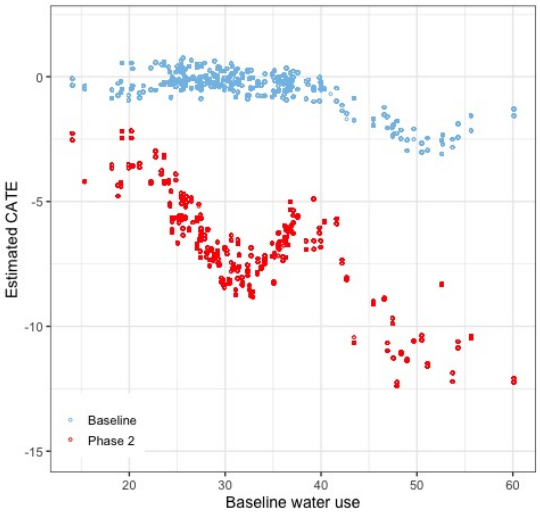
(c) *18L goal group in Phase 1*



(d) *18L goal group in Phase 2*



(e) *28L goal group in Phase 1*



(f) *28L goal group in Phase 2*

Notes. Each figure shows the estimated CATE for each treatment group $g \in \{MS, 18L, 28L\}$ in each phase $t \in \{1, 2\}$ using the cluster-robust causal forest algorithm of Athey et al. (2019).

to loss aversion or fixed penalty around the goals (both yield similar predictions), or possibly from psychological disengagement when goals are too difficult. For loss aversion or fixed penalty, the outcome would depend on which direction the goal is adjusted from the initial hard goal. By contrast, for psychological disengagement, it would not matter which direction the initial hard goal is adjusted as the subjects simply stop paying attention to the goal. We are unable to distinguish between the underlying mechanisms with our experimental design since we did not further increase the difficulty of the initial hard goal in phase 2. A promising avenue for future research is to tease apart the competing mechanisms to explain how goal-setting affects behavior in a dynamic setting as ours.

Our findings have implications for policymakers and management since they underscore the importance of goal-setting on effort provision and performance. We provide the first empirical evidence of the insidious effects of setting too hard an initial goal on subsequent effort provision. The level of goal difficulty should be chosen at an appropriate level, keeping in mind that it would not only affect task performance in the current period, but potentially in future periods. The caveat is that setting a goal that is too hard may not only lead to sub-optimal performance in the near term, but could result in permanent, diminished effects that cannot be easily undone by simply changing the goal. An important next step would be to examine if such dynamics of goal-setting extend to other domains beyond resource conservation.

In addition, we highlight the efficacy of goal-setting, coupled with real-time feedback, in settings where marginal costs to individuals are low (or zero, in our setting), as is often the case under certain rental agreements. While Myers and Souza (2020) find that behavioral channels such as competitiveness, social norms, or moral suasion combined with home energy reports, fail to increase energy-saving efforts in the absence of monetary incentives, our findings suggest otherwise. In particular, goal-setting with real-time feedback can serve as a powerful behavioral tool for promoting resource conservation, even without pecuniary motivations. Our work is in line with Tiefenbeck et al. (2019), who show that real-time feedback (on energy use) in the shower leads to higher energy-savings per shower among hotel guests, in a similar setting with no monetary incentives. This provides policymakers with a new set of non-price interventions to promote resource conservation, which goes beyond the traditional pecuniary approach of conservation taxes, rebate programs and subsidies.

References

Abrahamse, W., Steg, L., Vlek, C., and Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, 25(3):273–291.

Abreu, D., Pearce, D., and Stacchetti, E. (1990). Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica: Journal of the Econometric Society*, pages 1041–1063.

Agarwal, S., Fang, X., Goette, L., Sing, T. F., Staake, T., Tiefenbeck, V., and Wang, D. (2018). The role of goals and real-time feedback in resource conservation: Evidence from a large-scale field experiment. Technical report, National University of Singapore.

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

Anderson, A. and Green, E. A. (2018). Personal bests as reference points. *Proceedings of the National Academy of Sciences*, 115(8):1772–1776.

Asmus, S., Karl, F., Mohnen, A., and Reinhart, G. (2015). The impact of goal-setting on worker performance-empirical evidence from a real-effort production experiment. *Procedia CIRP*, 26:127–132.

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

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51.

Attari, S. Z., DeKay, M. L., Davidson, C. I., and De Bruin, W. B. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of sciences*, 107(37):16054–16059.

Avoyan, A., Khubulashvili, R., and Mekerishvili, G. (2020). Call it a day: History dependent stopping behavior.

Becker, L. J. (1978). Joint effect of feedback and goal setting on performance: A field study of residential energy conservation. *Journal of Applied Psychology*, 63(4):428–433.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brookins, P., Goerg, S., and Kube, S. (2017). Self-chosen goals, incentives, and effort. *Unpublished manuscript*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *The American Economic Review*, 99(4):1145–1177.

Clark, D., Gill, D., Prowse, V., and Rush, M. (2020). Using goals to motivate college students: Theory and evidence from field experiments. *Review of Economics and Statistics*, 102(4):648–663.

Corgnet, B., Gómez-Miñambres, J., and Hernán-Gonzalez, R. (2015). Goal setting and monetary incentives: When large stakes are not enough. *Management Science*, 61(12):2926–2944.

Drucker, P. F. (1954). *The Practice of Management*. Harper, Reissue, Edition 2006.

Erez, M. (1977). Feedback: A necessary condition for the goal setting-performance relationship. *Journal of Applied Psychology*, 62(5):624.

Erez, M. and Zidon, I. (1984). Effect of goal acceptance on the relationship of goal difficulty to performance. *Journal of applied psychology*, 69(1):69.

Fang, X., Goette, L., Rockenbach, B., Sutter, M., Tiefenbeck, V., Schoeb, S., Staake, T., et al. (2020). Complementarities in behavioral interventions: Evidence from a field experiment on energy conservation. *CRC TR 224 Discussion Paper*.

Ferraro, P. J. and Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.

Fisher, G., Kotha, S., and Lahiri, A. (2016). Changing with the times: An integrated view of identity, legitimacy, and new venture life cycles. *Academy of Management Review*, 41(3):383–409.

Goerg, S. J., Kube, S., and Radbruch, J. (2019). The effectiveness of incentive schemes in the presence of implicit effort costs. *Management Science*, 65(9):4063–4078.

Goette, L., Leong, C., and Qian, N. (2019). Motivating household water conservation: A field experiment in Singapore. *PloS one*, 14(3).

Grove, A. S. (1983). *High Output Management*. Vintage; 2nd Edition (August 29, 1995).

Haenni, S. (2019). Ever tried. ever failed. no matter? on the demotivational effect of losing in repeated competitions. *Games and Economic Behavior*, 115:346–362.

Harding, M. and Hsiaw, A. (2014). Goal setting and energy conservation. *Journal of Economic Behavior & Organization*, 107:209–227.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Heath, C., Larrick, R., and Wu, G. (1999). Goals as Reference Points. *Cognitive Psychology*, 38:79–107.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–291.

Kennerley, M. and Neely, A. (2003). Measuring performance in a changing business environment. *International Journal of Operations & Production Management*.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Langenbach, B. P., Berger, S., Baumgartner, T., and Knoch, D. (2019). Cognitive resources moderate the relationship between pro-environmental attitudes and green behavior. *Environment and Behavior*, page 0013916519843127.

Latham, G. P. and Locke, E. A. (2006). Enhancing the benefits and overcoming the pitfalls of goal setting. *Organizational Dynamics*, 35(4):332–340.

Levin, J. (2003). Relational incentive contracts. *American Economic Review*, 93(3):835–857.

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational behavior and human performance*, 3(2):157–189.

Locke, E. A. and Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705.

Locke, E. A., Shaw, K. N., Saari, L. M., and Latham, G. P. (1981). Goal setting and task performance: 1969–1980. *Psychological bulletin*, 90(1):125.

Myers, E. and Souza, M. (2020). Social comparison nudges without monetary incentives: Evidence from home energy reports. *Journal of Environmental Economics and Management*, page 102315.

Ordóñez, L. D., Schweitzer, M. E., Galinsky, A. D., and Bazerman, M. H. (2009). Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives*, 23(1):6–16.

Oyer, P. (2000). A theory of sales quotas with limited liability and rent sharing. *Journal of Labor Economics*, 18(3):405–426.

Pritchard, R. D. and Curts, M. I. (1973). The influence of goal setting and financial incentives on task performance. *Organizational Behavior and Human Performance*, 10(2):175–183.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. (2018). Overcoming salience bias: how real-time feedback fosters resource conservation. *Management Science*, 64(3):1458–1476.

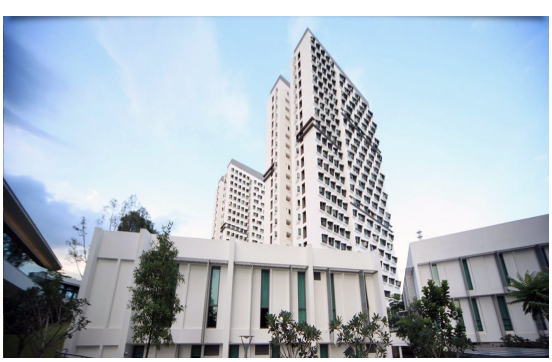
Tiefenbeck, V., Wörner, A., Schöb, S., Fleisch, E., and Staake, T. (2019). Real-time feedback promotes energy conservation in the absence of volunteer selection bias and monetary incentives. *Nature Energy*, 4(1):35–41.

Tversky, A. and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics*, 106(4):1039–1061.

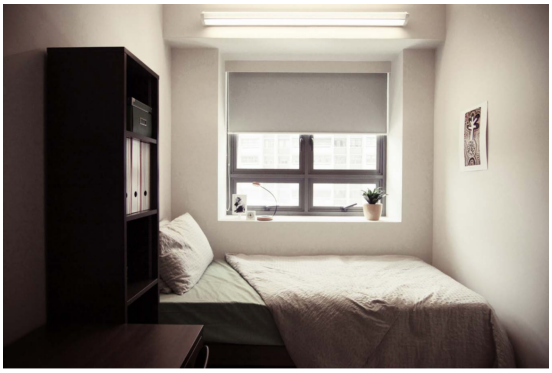
Appendix

A. Supplemental analyses and material

Figure A1: *Experimental site*



(a) *Tembusu and Cinnamon colleges*



(b) *Single room*

(c) *Bathroom with 2 shower facilties*

Notes. The figure shows photographs of the residential colleges where we implemented the field experiment.

Figure A2: *Sample poster by treatment group*

(a) *MS group in Phase 1*

(b) *MS group in Phase 2*

(c) *18L goal group in Phase 1*

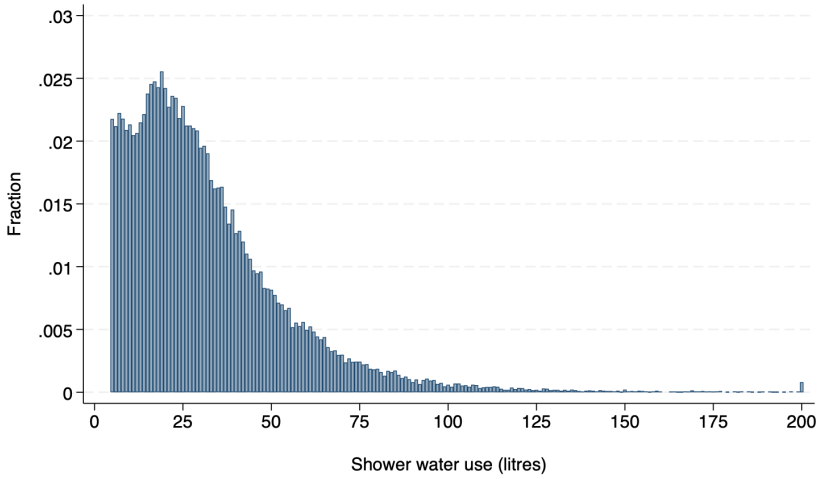
(d) *18L goal group in Phase 2*

(e) *28L goal group in Phase 1*

(f) *28L goal group in Phase 2*

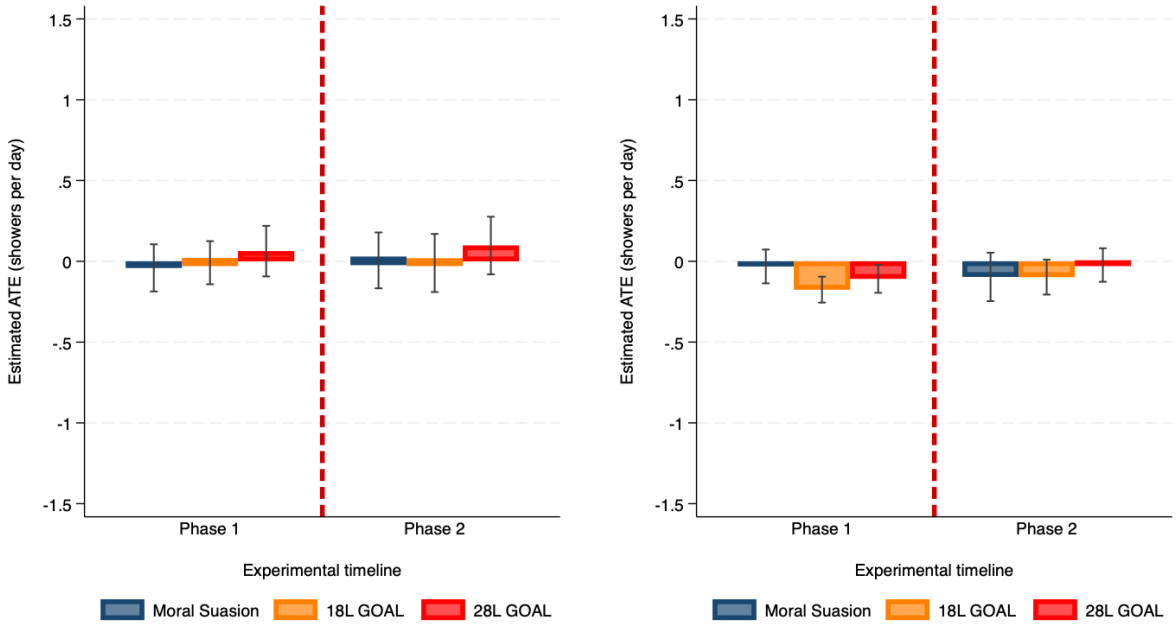
Notes. A specific type of poster was displayed in each shower facility based on the assigned experimental condition in each phase.

Figure A3: *Distribution of baseline shower water use*



Notes. The figure shows the histogram of water use per shower of all live showers in the baseline period. To reduce the influence of outliers, we exclude observations that recorded under 4 litres (inclusive) of water use and top-code values exceeding 200 litres, as these instances are unlikely to be actual shower events.

Figure A4: *Estimated ATEs (extensive margin) for above and below median users*



(a) *Above-median users*

(b) *Below-median users*

Notes. Each bar represents the difference-in-differences estimates of number of showers per day for the experimental groups in phase 1 and 2, respectively, relative to the control group. Panel (a) displays the estimated ATEs for the sample of above-median users, while panel (b) displays the estimated ATEs for the sample of below-median users. The error whiskers display +/− one standard error of the mean. To reduce the influence of outliers, we exclude observations that recorded under 4 litres (inclusive) of water use, as these instances are unlikely to be actual shower events.