Discussion Paper Series – CRC TR 224

Discussion Paper No. 283
Project B 07

The Dynamics of Goal Setting: Evidence From a Field Experiment on Resource Conservation

Lorenz Goette [1]
Hua-Jing Han [2]
Zhi Hao Lim [3]

April 2021

[1] University of Bonn and National University of Singapore. Adenauerallee 24-42, 53113 Bonn, Germany. 1 Arts Link AS2, Singapore. Email: lorenz.goette@uni-bonn.de
[2] University of Bonn. Adenauerallee 24-42, 53113 Bonn, Germany. Email: han@uni-bonn.de
[3] National University of Singapore. 1 Arts Link AS2, Singapore. Email: ecslzh@nus.edu.sg

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

The Dynamics of Goal Setting: Evidence from a Field Experiment on Resource Conservation*

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

November 16, 2020

Abstract

In a field experiment at two residential colleges, we use moral suasion and real-time feedback to examine how residents respond to different goals on shower water use over time. In phase 1 of our intervention, we find significant conservation effects for residents assigned to either the hard or moderate goal, but surprisingly detect no significant differences between both groups. When the goals are adjusted to a common, intermediate level in phase 2, we see a divergence in average treatment effects, driven by underperformance of the group with the initial hard goal. Interestingly, we find that high baseline users in the moderate goal group display larger conservation effects throughout the intervention, but this pattern is not observed in their counterparts with the initial hard goal. Our results suggest that setting too hard a goal can permanently diminish motivation, and adjusting the goal does not undo the initial damage.

JEL classification: C93, Q41, D91

Keywords: field experiment, goal setting, resource conservation

[*] Our research assistant team in Singapore provided invaluable support in implementing the experiment. We thank Ximeng Fang, Sebastian Kube and Johannes Weber for helpful comments. We would also like to acknowledge the generous financial support from 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.

[†] University of Bonn and National University of Singapore. Adenauerallee 24-42, 53113 Bonn, Germany. 1 Arts Link AS2, Singapore. Email: lorenz.goette@uni-bonn.de.

[‡] University of Bonn. Adenauerallee 24-42, 53113 Bonn, Germany. Email: han@uni-bonn.de.

[§] National University of Singapore. 1 Arts Link AS2, Singapore. Email: ecslzh@nus.edu.sg.

1 Introduction

Goals are widely used to motivate individuals. In the private and public sectors, goals (or objectives) are viewed as a key instrument to manage motivation and effort provision among employees (see, e.g. Drucker, 1954; Grove, 1983).[1] In economics, several strands of literature have examined how contracts that feature a discrete goal, with a bonus attached to it, can emerge as optimal incentive schemes in the presence of moral hazard problems.[2]

A large literature in psychology examines how goals also appear to affect motivation above and beyond the economic incentives that they may be coupled with. In general, a goal works well given commitment and attainability of the goal, and if there are no conflicting goals. In addition, difficult goals appear to have a higher motivating effect than easy goals (see, e.g. Locke and Latham, 1990, 2002, 2006). Heath et al. (1999) develop an interpretation that is particularly relevant to economics: in a series of hypothetical scenarios, they show that goals seem to inherit the properties of reference points in prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991).[3] Individuals behave towards goals in a manner consistent with experiencing loss aversion and diminishing sensitivity around them.[4] [5]

Goals may change over time. In management, business objectives may change due to economic shifts or evolution of a firm (see, e.g. Fisher et al., 2016; Kennerley and Neely, 2003). If goals directly affect individuals' motivation, changing them poses additional challenges. For instance, consider two goals, M and H, where M is a moderate goal which is easier to reach than H which is a hard goal. Consider two individuals, one of which starts with the moderate goal M and the other with the hard goal H. If goals

[1] In particular, a goal setting approach called Objectives and Key Results (OKRs) is enjoying great popularity of late. Attributed to the former Intel CEO Andrew S. Grove, and popularized by Doerr (2018), OKRs are adopted by many successful Silicon Valley companies, including Google, Oracle and Dropbox.

[2] Oyer (2000) shows how such schemes can be optimal in the presence of limited-liability constraint. Levin (2003) examines how such contracts can arise as efficient equilibrium of relational contracts in a repeated-game setting. More generally, Abreu et al. (1990) show that in repeated games with imperfect monitoring, so-called bang-bang equilibria, which could be interpreted as a goal with a bonus, can emerge.

[3] 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.

[4] Allen et al. (2017) further expand on the concept of goals as reference points by showing that there are settings where goals as reference points are not rational expectations and not necessarily linked to the status quo.

[5] Herweg et al. (2010) show that, with loss averse individuals, goal contracts naturally emerge as optimal in the Kőszegi and Rabin (2006) framework. The reason is that a binary goal provides the best trade-off between incentives and and the required compensation for sensations of losses that the chosen contract induces.

serve as reference points, it may well be that individuals exert more effort to reach M, while diminishing sensitivity leads them to exert less effort to try and reach H (Heath et al., 1999). Consider now moving both individuals to an intermediate goal I, which is halfway between M and H. For individuals starting out at M, the goal has become harder. However, because their new reference point is (near) M, loss aversion incentivizes them to work hard in order to try and reach I. By contrast, individuals who started out with the hard goal H now feel that the new goal is easier relative to their reference point. Working towards it feels like a gain, and consequently, effort will be lower. Thus, even though all individuals now face the same goal I, those who started out with the moderate goal M perform better in the long run than those who started out with the hard goal H.

In this paper, we conduct a field experiment to test this central prediction. The experiment is set in the context of household water conservation by individuals. Our sample consists of over 600 students in two residential colleges at the National University of Singapore. It is important to note that residents are charged a flat fee for their water (and energy) use. Thus, the setting allows us to focus on the effects of goals on non-pecuniary motivations.

We use moral suasion and real-time feedback to analyze the effects of changing goals on shower water use. More specifically, we seek to answer two questions of interest: First, how does the degree of difficulty of the initial goal affect performance? Second, when the goals are subsequently adjusted to a common, intermediate level, does it improve or dampen previous conservation efforts?

In phase 1 of the experiment, we assigned subjects to one of four groups. In our two key experimental groups, subjects are encouraged to keep their water use per shower below a specified target: one condition received an 18L goal, the other condition received a 28L goal. Baseline shower water use was around 32L on average, thus making the former a hard goal to reach (requiring halving of the water use to reach it), and the latter a moderate goal (requiring to save only about 15 percent of water per shower). Smart shower heads provided real-time feedback relative to the goal: They shine in a green light at the beginning of a shower, and subsequently change color to yellow, followed by orange and then red with increasing water use. Finally, the shower heads display a blinking red light when the water volume exceeds the specified shower goal. These thresholds were clearly communicated through posters in each shower facility. We name the two key experimental groups as *18L GOAL* and *28L GOAL* groups, respectively. To control for potential moral suasion, a third experimental group only received a poster encouraging users to keep water use below 28L, but no feedback to track their behavior. Finally, the fourth experimental group (control) received neither real-time feedback nor moral suasion. In phase 2 of the experiment, we retained the above setup, but changed the initial goals (either 18L or 28L) to a common, intermediate goal of 24L.

Overall, we find that moral suasion alone did not induce any significant effect on water use. However, the provision of real-time feedback relative to a goal led to large and significant conservation effects of around 15% of baseline shower water use. Interestingly, we do not find significant differences in the average treatment effects between the *18L GOAL* and the *28L GOAL* groups in phase 1. Notwithstanding, when both groups are assigned the intermediate goal (24L) in phase 2, we observe a divergence in average treatment effects: subjects assigned the initial hard goal (18L) systematically underperform under the 24L goal relative to their counterparts who were assigned the initial moderate goal (28L). In addition, we find evidence of heterogeneity (masked by similar average treatment effects in phase 1): high baseline users in the *28L GOAL* group show larger conservation effects than their counterparts in the *18L GOAL* group. Further, this heterogeneity persists in phase 2, even though both groups now receive the same goal.

Thus, in phase 1 of our experiment, we show, as e.g. Goerg et al. (2019) and Agarwal et al. (2018) documented, that harder goals do not necessarily lead to better outcomes. As discussed in Agarwal et al. (2018), we also find differential interaction effects with baseline water use for our *18L GOAL* and *28L GOAL* groups. However, we additionally show that these heterogeneous effects carry over to the next phase, even when both goals are changed to the same level. We are, to our knowledge, the first to show that setting too hard a goal is associated with lasting, detrimental effects on effort and performance, even after the goal had been adjusted to an intermediate level. The initial goal assignment is thus critical and leads to long-term effects on performance, which cannot be easily undone. In our setting, this effect was particularly pronounced for high baseline users in the *18L GOAL* group: the hard goal muted their conservation efforts right from the outset, and this effect persisted throughout the intervention.

Our work is related to two main strands of literature. First, it contributes to the literature on the role of goal setting on motivation and effort provision. In their widely acknowledged paper, Locke and Latham (2002) describe a positive, linear function between goal difficulty and effort and performance. In particular, Locke and Latham (2006) specify that difficult goals have a stronger positive impact on task performance than easy goals, given commitment, attainability and the absence of conflicting goals. In addition, Erez (1977) shows that feedback is a necessary condition for the positive goal-performance relationship. Notwithstanding, Ordóñez et al. (2009) describe goals as having "powerful and predictable side effects", and state that goal-setting should come with a "warning label". A slew of measures have been proposed to account for the possible pitfalls of goal-setting, e.g. Latham and Locke (2006) list ten possible pitfalls of goal-setting, while Ordóñez et al. (2009) recommend a more cautious approach to goal-setting and include a list of questions for managers to consider when setting goals.

But none of these measures specifically addresses the potential pitfalls that might arise when goals change over time, except for the fact that in business environments which might be changing, performance goals might actually prevent learning. Locke and Latham (2002) propose the use of learning goals instead of performance goals in "complex environments", but this may not always be feasible outside business scenarios. The economics literature is subdivided into literature on exogenous and endogenous goals. Endogenous goals are self-chosen goals (see e.g. Brookins et al. (2017), who provide a good overview of the existing literature on endogenous goals). Harding and Hsiaw (2014) find that for self-set goals on resource conservation, realistic goals will lead to more savings than very low or unrealistic high goals.[6] Our paper falls under the category of exogenous goals, which are set by another party. In general, exogenous goals have a positive effect on performance, in particular when they are attainable by agents (Corgnet et al., 2015; Gómez-Miñambres, 2012; Wu et al., 2008). However, to our knowledge, the economic literature on the goal-performance relationship for exogenous goals is mostly theoretical or empirically studied in the lab. We add to these findings with a setting that allows us to look at choices without monetary incentives. Moreover, in the economic literature, goals can be seen as reference points as mentioned above (Heath et al., 1999; Kahneman and Tversky, 1979; Tversky and Kahneman, 1991). If goals can have the properties of reference points, individuals might experience related effects such as loss aversion and diminishing sensitivity around

[6] Their setting did not include real-time feedback. Participants would receive monthly feedback (and bonus points) on their energy use. However, this feedback was not directly related to their goals. Additionally, they could monitor their energy use via a website which the average consumer would log in to every 2-3 months.

them which can pose challenges when goals are changed. Notwithstanding, the literature on the effect of changing goals is sparse. Our contribution to the literature on goal-setting is twofold: First, by examining the goal-performance relationship for exogenous goals in a randomized field experiment. Second, by looking at "dynamic goals", i.e. how the goal-performance relationship plays out once goals are changed. In particular, we test whether individuals who start with a moderate goal will perform better once goals are changed to the same intermediate goal, compared to individuals who start with a hard goal.

Second, our paper is related to the growing literature on the efficacy of behavioral tools for resource conservation, specifically in the context of water (or energy) use in the shower. Recent research highlights that limited attention and imperfect information by households play an important role in shaping resource consumption (Attari et al., 2010; Chetty et al., 2009; Langenbach et al., 2019; Tiefenbeck et al., 2018).

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)). We contribute to this strand of literature by carefully examining the complementarities of goals and real-time feedback in a randomized field experiment using smart shower heads as tools for resource conservation, with an emphasis on the effects of initial goal difficulty. Our unique setting further allows us to test these behavioral interventions in the absence of monetary incentives.

The rest of the paper is organized as follows. Section 2 describes the experimental design and outlines the behavioral predictions of our treatments. Section 3 presents the descriptive evidence and formal analysis. Section 4 provides robustness checks to rule out alternative explanations. Section 5 concludes.

2 Experimental Design

We conducted a randomized field experiment at two neighboring residential colleges ("Cinnamon" and "Tembusu") in University Town of the National University of Singapore, from August 5, 2019 to November 24, 2019. This was in partnership with the NUS Office of Housing Services, which was keen on exploring behavioral interventions to promote resource conservation on campus.

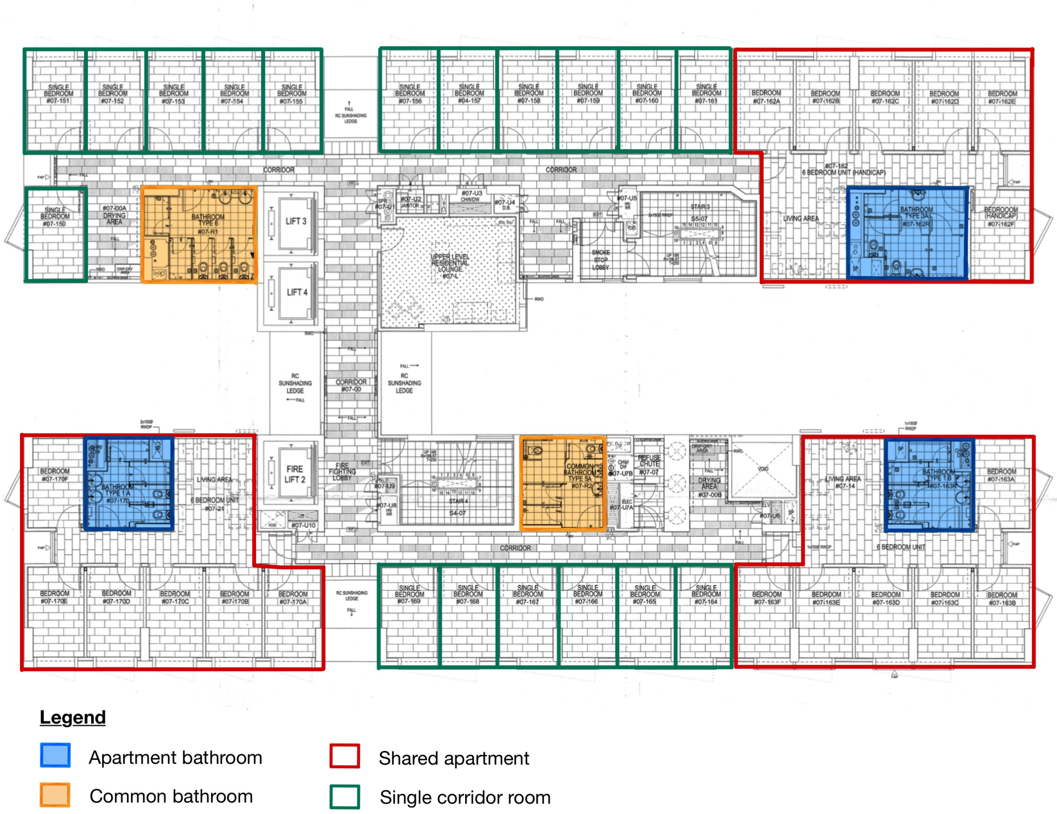
2.1 Background

Each residential college consists of 21 storeys with over 600 rooms in total, providing accommodation to local undergraduates, international exchange students, and a small grp of faculty members. At the beginning of each semester, students can opt for their preferred room type (i.e. single corridor room or single room in shared apartment) on either mixed or single-gender floors. Our pool of subjects comprises mainly incoming freshmen and excludes all faculty members. Figure A1 displays some photos of the experimental site at Cinnamon and Tembusu colleges.

In total, 324 HYDRAO smart shower heads were installed in all designated bathrooms at Cinnamon and Tembusu colleges. Note that there are two types of bathrooms on each floor: apartment and common bathrooms (refer to Figure 1). Residents who live in a shared apartment have access to their own apartment bathroom, while those who stay in the single corridor rooms use the common bathrooms.[7] 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.

[7] Note that a resident who stays in a single corridor room would not have access to the apartment bathroom.



Notes. The figure shows the floor plan that is representative 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.

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

2.2 HYDRAO smart shower head

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 the level of water use. This provides users with real-time feedback about their shower water use. The exact thresholds and corresponding colors can be configured, which allows us to implement different (exogenous) goals for our treatment groups.

For each shower event, the smart shower head can collect real-time data when it is connected to the server via WiFi. As a safeguard, there is an internal memory of 200 shower events. This means that if a particular shower fails to be transmitted real-time, the data will be stored and transmitted as an offline shower event as soon as the WiFi connection is re-established. If the shower is interrupted for a short duration of time (e.g.

for soaping purposes), the smart shower head will still consider it the same shower event as long as the interruption is under 2 minutes. Beyond 2 minutes, the shower head will assume that a new shower event has started.

Additionally, the smart shower head does not require external power supply as it is powered by water flow through a mini-turbine. For home usage, there is a HYDRAO shower app which can be synchronized with the shower head to configure the color thresholds. For the purpose of r experiment, we remotely set the thresholds for all shower heads with the use of gateways, so as to ensure minimal disruption to the residents. Importantly, our subjects were not informed of the app and could not change the configured settings of the shower heads from their end. This protects the integrity of our randomisation.

2.3 Treatment assignment

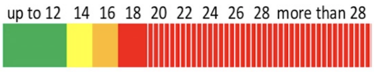
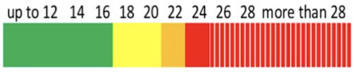
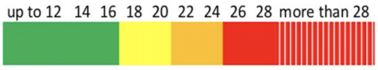
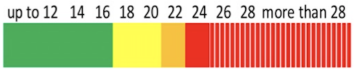
Our experiment comprises three stages: baseline, phase 1 and phase 2. The baseline period was in effect for 6 weeks from the beginning of the semester (i.e. August 5, 2019 to September 15, 2019). Phase 1 of the intevention corresponds to the next 5 weeks, which took place from September 16, 2019 to October 21, 2019. We then transitioned to phase 2 for the rest of the semester (i.e. October 22, 2019 to November 24, 2019).

In the baseline period, no intervention was implemented in all shower facilities. The primary objective was to collect information on pre-experimental showering behavior of the residents. Our baseline data contain observable characteristics from each device (used interchangeably with shower head), such as water use per shower, number of showers per day and flow rate. We use this information to conduct randomization checks in the following section.

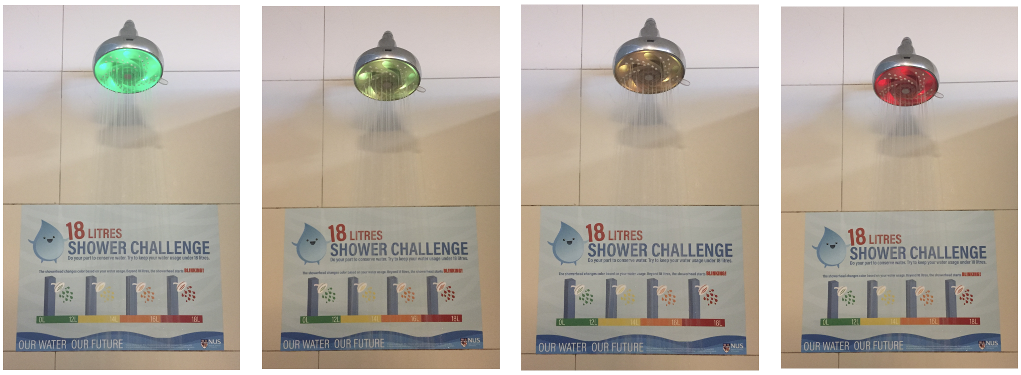
For the intervention, we implemented four experimental groups: a control and three treatments. These assignments were permanent throughout the rest of the experiment. The *Control* group received neither the shower poster nor real-time feedback in both phases. The *Moral Suasion (MS)* group received a shower poster appealing to users to keep their water use under a specified level, but no feedback through the shower heads. The shower poster references a goal of 28L in phase 1, and subsequently 24L in phase 2. The *18L GOAL* group received a shower poster and real-time feedback that corresponds to the goal of 18L in phase 1, and thereafter 24L in phase 2. The *28L GOAL* group received a shower poster and real-time feedback that corresponds to the goal of 28L, which was similarly switched to 24L in phase 2. In sum, the treatment groups received different (exogenous) goals in phase 1 but identical goal in phase 2, with the information conveyed through the respective shower posters (see Figures A2, A3 and A4).

For the *18L GOAL* and *28L GOAL* groups, we programmed the shower heads to display real-time feedback, in the form of colored lights (resembling the traffic light system) that correspond to a set of thresholds. The shower head would display a green light at the start of each shower, then progress to yellow, orange and red light with increasing water consumption. When the volume of water use exceeds the goal, the shower head would begin to display blinking red light. Table 1 summarises 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.

Table 1: *Summary of treatment assignments*

Stage **Group**	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.



(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. assigned goal in phase 1), the shower head starts to display blinking red light.

Figure 2: *Implementation of 18L GOAL group*

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 (exogenous) goals. In addition, the implementation of different goals in phase 1 but the same goal in phase 2 allows us to study the relationship between goal difficulty and effort provision in a dynamic setting. In phase 1, we can compare the short-run effect of a moderate goal (28L) relative to a difficult goal (18L) on shower water use. In phase 2, we can then examine whether the initial goals have any bearing on how subjects respond to the new, intermediate goal (24L).

2.4 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; Kőszegi and Rabin, 2006).

If goals act as reference points, then the psychological utility of reaching the goal depends on loss aversion and diminishing sensitivity. 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 ambitious, they may elicit lower effort from individuals because of diminishing sensitivity relative to the reference point (Heath et al., 1999).

In our experiment, we chose the 18L and 28L goal such that they allow us to test this prediction in phase 1 of our experiment: the 28L goal was set such that it was attainable with reasonable effort, while the 18L goal was set to be extremely difficult to meet given baseline shower behavior.[8] We thus arrive at the following prediction:

Hypothesis 1: ***Initial goal difficulty and effort.*** *In phase 1, conservation effects in the 28L condition are greater than in the 18L condition.*

Several different mechanisms of how goals affect reference points could lead to hypothesis 1: it could 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. The main features we are interested in testing is how individuals respond when the difficulty of goals is changed over time. For the 28L condition, phase 2 of the experiment introduces a tougher goal of 24L. How the two groups respond to introducing 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 for the groups who formerly had the 18L and 28L goal, respectively, should lead to the same reference point, and hence to the same outcomes. We summarize this as

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

By contrast, if recent expectations or lagged outcomes shape reference points, then changing goals over time could have different effects: moving from the 18L goal to the 24L goal makes the goal easier relative to the previous benchmark, thus potentially moving individuals into the gain domain and thus reducing the marginal benefit from conserving water and hence conservation efforts. By contrast, having settled on the 28L goal, the shift to the new 24L goal represents a new tightening of the goal, thus pushing individuals again into the loss domain. Thus, switching to the 24L goal would generally reduce motivation to conserve in the 18L group, while increasing it in the 28L group. We summarize this in

[8] The average shower water use in the baseline period is 31.9 liters, with standard deviation of 23.0. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters.

Hypothesis 2b: ***Lagged expectations of outcomes affect reference points.*** *If reference points are affected by lagged outcomes or recent expectations, the 28L group will outperform the 18L group in conservation efforts in Phase 2.*

Finally, we test whether reference dependence makes gradual tightening of goals more effective. This stems from the same reasoning as above: The initial goal of 28L encourages conservation because the goal creates an immediate loss (relative to the baseline consumption level), but is not too difficult to reach. Thus, the marginal benefit of conservation efforts is high. As the goal can be reached, the new outcome sinks in as the reference point (or the expectation of the outcome). Once that happened, the new goal of 24L now again creates sensations of loss and raise the marginal benefit of conservation, thus lowering resource use even further. We summarize this as

Hypothesis 3: ***Gradual tightening of goals.*** *Switching the 28L group to the 24L goal leads to a significant increase in conservation efforts.*

Note that the difference between hypothesis 2b and hypothesis 3 is that hypothesis 2b compares the relative efforts between the two treatment groups (18L and 28L) in phase 2, while hypothesis 3 is about the change in the performance of the 28L group relative to the control group in phase 2.

The mechanism derived from changes in reference outcomes leads to similar predictions as when there is habit formation in the sense of Stigler and Becker (1977): In their model, such an effect could result from habit stock building up over time, and thus gradually reducing the marginal utility of showering and hence increasing conservation efforts over time.[9]

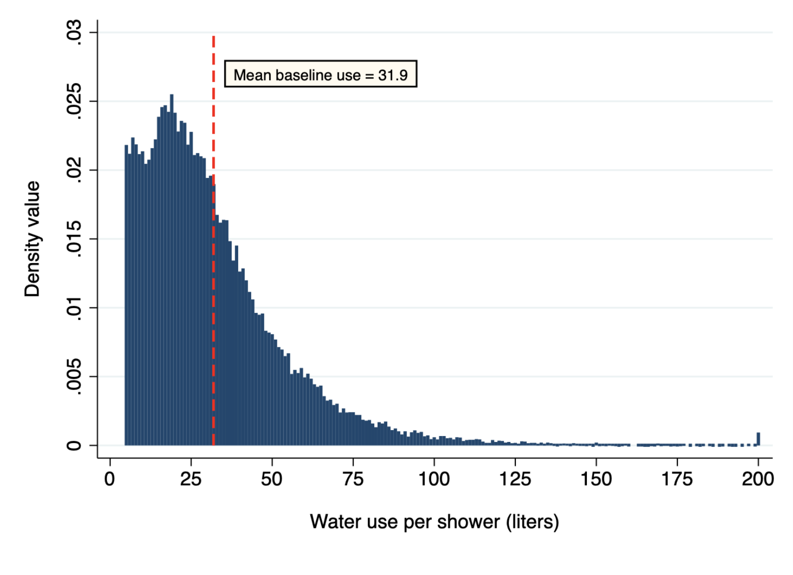
3 Analysis and Interpretation

Our data comes from the randomized field experiment described above. In total, we recorded 128,323 shower instances (of strictly more than 4 liters), from 301 working shower heads. The observations can be classified into two types: live showers and offline showers. The former refer to shower instances where data was transmitted real-time (at the point of showering) from the shower heads to our server — this gives us information about

[9] Byrne et al. (2018) test the consumption habit model explicitly in a similar setup.

the actual date and time each shower event took place. In contrast, the latter refer to shower instances that were transmitted with a time lag, and so we are unable to accurately pinpoint the occurrence of each shower event. To increase precision of our estimates, we thus consider the sample of 116,891 live showers (instead of all recorded showers) as our primary data source for analysis.

In the baseline period, we record a stable pattern of around 1,300 live showers on a regular weekday, and about half the number on weekends. On the intensive margin, we observe a right-skewed distribution of baseline water use per shower, with a mean of 31.9 liters (see Figure 3).



Notes. The figure shows the histogram of water use per shower using the sample of live showers in the baseline period. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters.

Figure 3: *Distribution of baseline water use per shower*

3.1 Randomization checks

To begin, we perform balance tests to support the integrity of the randomization. Table 2 presents a comparison across different treatment groups, relative to the control, on the baseline averages of key observables. It is apparent that balance of treatment is attained as almost all observables, in particular water use per shower and fraction of live showers, do not differ across groups. There is only slight statistical difference in the number of days since last transmission between the control and *28L GOAL* group.[10] This is largely

[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.

driven by a single shower head in the control group which rarely records live shower events, and thus does not constitute a cause for concern.[11] We conclude that our experimental groups are well-balanced and interpret any observed differences during intervention as causal treatment effects.

Table 2: *Randomization checks*

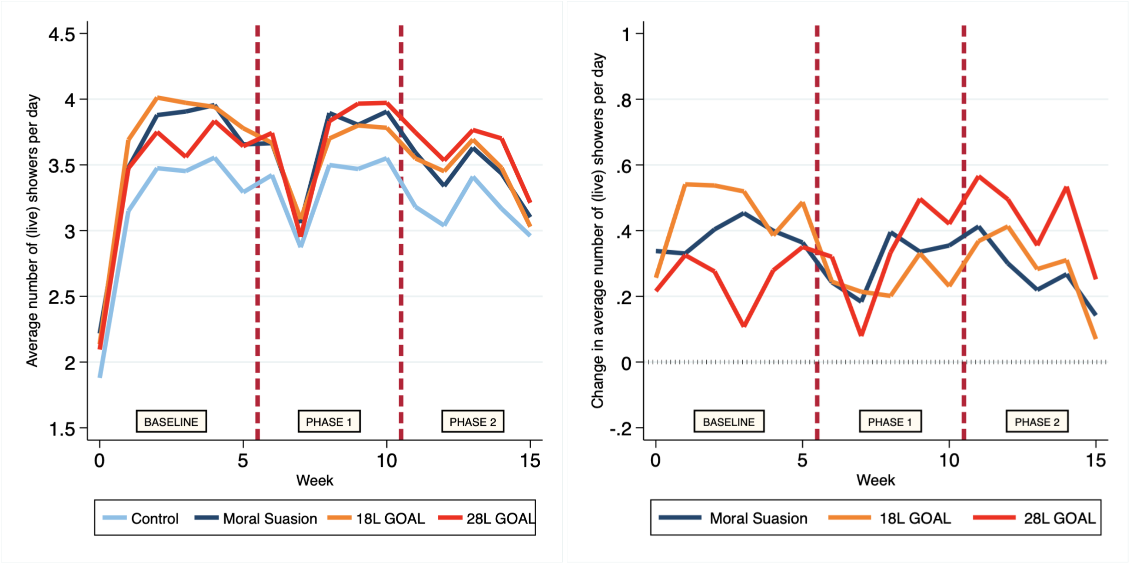
	Baseline averages by shower head						
Dependent variable:	Water use per shower (in liters)	Number of showers	Duration per shower (in seconds)	Fraction of *live* showers	Days since last transmission	Suite bathroom	Floor
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Moral Suasion	– 1.532	16.252	– 22.497	0.004	– 0.518	– 0.100	0.513
	(1.753)	(13.298)	(17.747)	(0.024)	(0.645)	(0.172)	(1.772)
18L GOAL	– 1.094	17.537	– 1.291	0.024	– 0.401	0.061	1.262
	(2.044)	(14.638)	(23.627)	(0.019)	(0.510)	(0.165)	(1.618)
28L GOAL	– 1.882	13.408	– 21.696	– 0.005	– 0.993**	– 0.024	– 0.194
	(1.626)	(12.811)	(16.982)	(0.033)	(0.460)	(0.171)	(1.985)
Constant	33.449***	135.414***	372.020***	0.883***	1.157**	0.586***	12.043***
	(1.249)	(8.480)	(13.727)	(0.017)	(0.455)	(0.121)	(1.293)
p-value for F-test	0.700	0.519	0.464	0.461	0.017	0.811	0.811
R^2	0.006	0.008	0.014	0.007	0.018	0.014	0.013
Observations	297	297	297	297	297	297	297

Notes. The results are obtained by regressing the various baseline averages of observables on assigned experimental groups. The omitted group is the control (i.e. received neither moral suasion nor real-time feedback). Standard errors clustered at the residence × floor × bathroom type level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.2 Descriptive evidence

In the context of goal-setting, our empirical analysis compares how moral suasion and real-time information feedback influence shower behavior on the extensive and intensive margins. We break the full sample into three distinct time periods: baseline, phase 1 and phase 2. Recall that in phase 1, the *18L GOAL* and *28L GOAL* groups received moral suasion and real-time feedback which referenced different goals (i.e. 18L vs. 28L), and subsequently in phase 2, both groups converged to the common goal of 24L. Figures 4 and 5 provide descriptive evidence of how our treatments impacted daily number of showers (extensive margin) and water use per shower (intensive margin) over time, respectively.

[11] In particular, the 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 experiment, only 11% are classified as live showers.



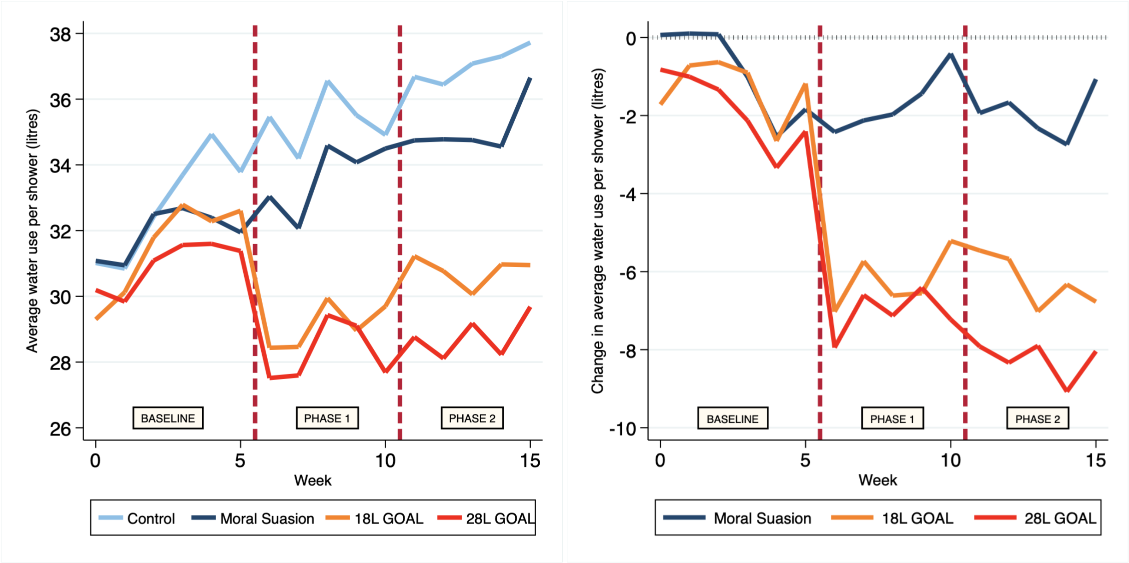
(a)

(b)

Notes. The daily number of *live* showers is averaged across all shower heads in the same experimental group on a weekly level. Panel (a) displays the daily average number of showers (per shower head) by experimental groups, and panel (b) displays the change in daily average number of showers, relative to the *control* group (i.e. received neither moral suasion nor real-time feedback). To reduce the influence of outliers, we do not consider observations which recorded under 4 liters (inclusive) of water use as shower instances.

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

On the extensive margin, we observe a stable pattern of around 3 to 4 daily showers (per shower head) across all groups over the course of the experiment. In particular, we do not observe any significant change in levels for any of the treatment groups during both phases of the intervention. It appears that our treatments have little to no effect on the extensive margin, which we will formally verify below. The only anomaly is a distinct drop in daily number of showers in week 7, but this is not a cause for concern as it coincides with the recess week, during which some residents may have left campus for a one-week break. In fact, it is reassuring that we witness the same drop in levels across all four experimental groups, which suggests that there is no differential selection (out of the experiment) during the mid-term break.



(a) **(b)**

Notes. Water use per shower (in liters) is averaged across all shower heads in the same experimental group on a weekly level. Panel (a) displays the average water use per shower by experimental groups, and panel (b) displays the change in average water use per shower, relative to the *control* group (i.e. received neither moral suasion nor real-time feedback). To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters.

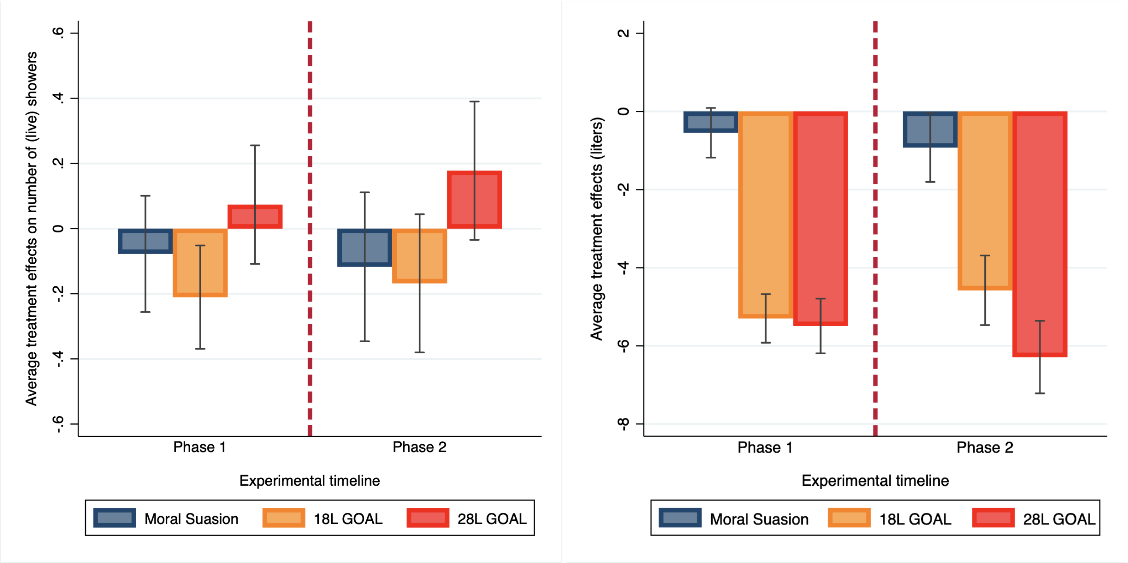
Figure 5: *Intensive margin of shower behavior by experimental groups*

On the intensive margin, we observe that all experimental groups have near-similar levels of mean water use per shower during the baseline period, consistent with our preceding randomization checks. For the control group, there is a visible upward trend of mean water use per shower over time; the *MS* group appears to fare slightly better in both phases of the intervention, but otherwise exhibits a similar upward trend. This stands in stark contrast with the *18L GOAL* and *28L GOAL* groups, which showed sharp reductions in mean water use per shower with the onset of real-time feedback (on top of moral suasion) in phase 1. We see that the large effects persist in phase 2, albeit to varying degrees depending on the initial assigned goal (18L vs. 28L). We will discuss this treatment effect dynamics in greater detail below.

To augment our analysis, Figure 6 displays the difference-in-differences estimates of the average treatment effects for the *MS*, *18L GOAL* and *28L GOAL* groups. We examine whether our treatments have an impact on the extensive and intensive margins of shower behavior in the left and right panels, respectively. The estimates are obtained by taking the difference between the outcome of interest in phase 1 and the baseline period, and similarly the difference between phase 2 and the baseline period.

First, it is clear from the left panel that there is no effect on the extensive margin (i.e. daily number of showers) in both phases of the intervention. This finding highlights that there is no differential selection into or out of the experiment across our treatment groups. By ruling out selection effects (changes in the composition of subjects), we can attribute any changes on the intensive margin as behavioral responses to our respective treatments.

Next, we turn to the right panel showing average treatment effects on the intensive margin (i.e. water use per shower). In phase 1, while there is only a modest decrease in mean water use in the *MS* group, we observe sharp reductions in the *18L GOAL* and *28L GOAL* groups that receive real-time feedback, in addition to moral suasion. The standard error bars around the means suggest highly significant effects. It also appears that both groups respond similarly to the treatment, despite receiving different goals (i.e. 18L vs. 28L). However, when both groups converged to the common 24L goal in phase 2, we observe a divergence in treatment effects. Again, we see that the use of moral suasion alone has only marginal effects, if any, on water use per shower.



(a) *Extensive margin*

(b) *Intensive margin*

Notes. Each bar represents the difference-in-differences estimates of the outcome of interest for each experimental group in phase 1 and 2 respectively, relative to the control group in the baseline period. Panel (a) focuses on the extensive margin by using number of showers as the outcome variable, while panel (b) looks at the intensive margin with water use per shower as the outcome variable. The error whiskers display +/− standard error of the mean. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. Equivalently, we do not consider observations which recorded under 4 liters (inclusive) of water use as shower instances.

Figure 6: *Average treatment effects by experimental groups*

3.3 Estimation strategy

To formally identify the respective treatment effects, we estimate the following model

$$\begin{aligned} y_{ith} = \alpha_i + \lambda_t + \gamma_h + \Big(\beta_{MS1}\mathrm{MS}_i + \beta_{18L1}18\mathrm{L}_i + \beta_{28L1}28\mathrm{L}_i\Big) \times \mathrm{PHASE1}_{ith} \\ + \Big(\beta_{MS2}\mathrm{MS}_i + \beta_{18L2}18\mathrm{L}_i + \beta_{28L2}28\mathrm{L}_i\Big) \times \mathrm{PHASE2}_{ith} + \epsilon_{ith} \end{aligned} \quad (1)$$

where y_{ith} is the outcome variable of interest, e.g. 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 a dummy 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 $18\mathrm{L}_i$ and $28\mathrm{L}_i$ variables are indicators for being assigned to the *18L GOAL* and *28L GOAL* groups respectively. Note that the MS_i variable is defined to be one for all three treatment groups, instead of only the *MS* group. This is because the *18L GOAL* and *28L GOAL* groups both receive moral suasion, and therefore the *MS* group serves as the relevant comparison for identifying the marginal effects of real-time feedback (on top of moral suasion). $\mathrm{PHASE1}_{ith}$ is a dummy variable that equals one for the period when the initial shower goals (i.e. either 18L or 28L) were introduced, whereas $\mathrm{PHASE2}_{ith}$ is a dummy variable that equals one for the latter period when the shower goal is changed to 24L.[12] $\epsilon_{i,t,h}$ is the random error term and standard errors are clustered at the residence × floor × bathroom type level (i.e. unit of randomization).

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 treatment conditions relative to the control. The average treatment effects are identified from within-device variation over time, controlling for aggregate hourly and daily shocks. To elaborate, β_{MS1} can be interpreted as the average treatment effect of moral suasion on water use per shower in phase 1 relative to the baseline, and β_{MS2} gives the corresponding average treatment

[12] To be specific, for the *MS* group, $\mathrm{PHASE1}_{ith}$ equals one from 4PM, September 16 to 5PM, October 22, whereas $\mathrm{PHASE2}_{ith}$ 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 $\mathrm{PHASE1}_{ith}$ and $\mathrm{PHASE2}_{ith}$ variables respectively.

effect in phase 2 relative to the baseline. Similarly, the coefficients on the interacted 18L_i and 28L_i variables represent the effect of real-time feedback over and above the effect of moral suasion in each respective phase. Table 3 presents the results.

First, it is evident from columns 1 and 2 that the use of moral suasion alone did not induce any effect on water use per shower in both phases. While the point estimates of between -0.7 and -0.8 run in the desired direction, they are statistically insignificant. Over and above the effect of moral suasion, the provision of real-time feedback induced large and significant conservation effects of around 15% (i.e. between 4.6 and 4.8 liters per shower), consistent with previous studies on water conservation that involves smart shower heads (Goette et al., 2019; Tiefenbeck et al., 2018). Interestingly, we observe that the implementation of different shower goals (i.e. 18L vs. 28L in phase 1) did not lead to any discernible difference in treatment effect of real-time feedback; we cannot reject the null hypothesis of equality between the 18L and 28L goals in phase 1 ($p = 0.783$). This runs counter to Hypothesis 1 on initial goal difficulty and effort, which predicts larger conservation effects in the 18L condition.

Next, when the shower goal is adjusted to 24L in phase 2, we observe a divergence of the average treatment effects. As shown in column 2, while the 18L condition fared worse, the 28L condition responded more strongly to real-time feedback in phase 2 relative to phase 1. In particular, we are able to marginally reject the null hypothesis of equality between the 18L and 28L conditions in phase 2 ($p = 0.061$). Comparing between the 18L and 28L conditions, we further test for equality of the change in treatment effects from phase 1 to 2, and can easily reject the null hypothesis at the 1% level ($p = 0.008$). Thus, our data soundly rejects Hypothesis 2a that the shower goals serve directly as reference points, as this interpretation would have yielded the same outcomes for both the *18L GOAL* and *28L GOAL* groups. On the flip side, our results lend credence to Hypothesis 2b that reference points are affected by recent expectations or lagged outcomes, as observed from the underperformance in conservation efforts by the *18L GOAL* relative to the *28L GOAL* group.

Finally, we consider Hypothesis 3 which states that the gradual tightening of goals for the *28L GOAL* group would lead to a significant increase in conservation efforts. While we see the effect size increasing for the 28L condition from phase 1 to 2, the point estimates are not significantly different ($p = 0.304$). Therefore, our data only provides suggestive evidence in support of the hypothesis at best.

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

Dependent variable:	Water use per shower (liters)	
	PHASE 1 (1)	PHASE 2 (2)
MS × PHASE	– 0.824 (0.512)	– 0.773 (0.859)
18L × PHASE	– 4.627*** (0.545)	– 3.785*** (0.864)
28L × PHASE	– 4.813*** (0.621)	– 5.427*** (0.803)
Constant	33.905*** (0.266)	
p-value for $\beta_{18L} = \beta_{28L}$	0.783	0.061
p-value for $\beta_{18L1} = \beta_{18L2}$	0.165	
p-value for $\beta_{28L1} = \beta_{28L2}$	0.304	
p-value for $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$	0.008	
p-value for F-test	0.000	
Device FEs	*Yes*	
Date FEs	*Yes*	
Hour-of-day FEs	*Yes*	
R^2	0.139	
Observations	116891	

Notes. This table shows the effects of moral suasion and real-time feedback (based off different shower goals in each phase) on water use per shower. 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 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. *live* showers) is reported in the last row. Standard errors clustered at the residence × floor × bathroom type level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Next, we examine how subjects adjusted their showering behavior on the extensive margin. Given that real-time feedback, coupled with moral suasion, induced large conservation effects of roughly 14% – 19%[13], it is pertinent to consider if the savings were offset by subjects taking more showers each day. On the flip side, if the treatments had induced subjects to take fewer showers each day, this might generate negative externalities in the

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

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

Table 4 presents the results, with all point estimates being statistically insignificant. This is consistent with the descriptive evidence presented above (see Figure 4). This allays our main concern about subjects compensating reduced water use per shower with greater frequency of showers each day. Notwithstanding, we are able to reject the null hypothesis of equal treatment effects between the 18L and 28L conditions in phase 1 at the 5% level ($p = 0.040$). This suggests that each shower head in the 28L condition registers 0.284 more showers per day, on average, relative to the 18L condition in phase 1. This is not a huge concern as each shower head registers an average of 3.41 live showers per day in the baseline period, so 0.284 showers (8%) constitute a relatively small fraction. Therefore, we conclude that our treatments only induced adjustments on the intensive margin, so we can focus our attention on it.

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

Dependent variable:	Number of showers per day	
	PHASE 1 (1)	PHASE 2 (2)
MS × PHASE	– 0.078 (0.197)	– 0.117 (0.245)
18L × PHASE	– 0.133 (0.117)	– 0.051 (0.199)
28L × PHASE	0.151 (0.150)	0.295 (0.224)
Constant	3.495*** (0.091)	
p-value for $\beta_{18L} = \beta_{28L}$	0.040	0.113
p-value for $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$	0.708	
p-value for $\beta_{MS} = \beta_{18L} = \beta_{28L} = 0$	0.172	0.419
p-value for F-test	0.410	
Device FEs	*Yes*	
Date FEs	*Yes*	
R^2	0.590	
Observations	33712	

Notes. This table shows the effects of moral suasion and real-time feedback (based off different shower goals in each phase) on number of *live* showers per day. 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 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations is reported in the last row. Standard errors clustered at the residence × floor × bathroom type level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.4 Underlying heterogeneity

Motivated by previous studies showing that high baseline users display larger conservation gains (Allcott, 2011; Ferraro and Price, 2013; Tiefenbeck et al., 2018), we seek to delve into the underlying behavioral mechanisms of our treatments. Formally, we augment our model by including interaction terms with mean baseline use of each shower head, $\bar{y_0}$.

$$\begin{aligned} y_{ith} = \alpha_i + \lambda_t + \gamma_h &+ \Big(\beta_{MS1}\mathrm{MS}_i + \beta_{18L1}18\mathrm{L}_i + \beta_{28L1}28\mathrm{L}_i\Big) \times \mathrm{PHASE1}_{ith} \\ &+ \Big(\beta_{MS2}\mathrm{MS}_i + \beta_{18L2}18\mathrm{L}_i + \beta_{28L2}28\mathrm{L}_i\Big) \times \mathrm{PHASE2}_{ith} \\ &+ \Big(\gamma_{MS1}\mathrm{MS}_i + \gamma_{18L1}18\mathrm{L}_i + \gamma_{28L1}28\mathrm{L}_i\Big) \times \mathrm{PHASE1}_{ith} \times \bar{y_0} \\ &+ \Big(\gamma_{MS2}\mathrm{MS}_i + \gamma_{18L2}18\mathrm{L}_i + \gamma_{28L2}28\mathrm{L}_i\Big) \times \mathrm{PHASE2}_{ith} \times \bar{y_0} \\ &+ \Big(\delta_1\mathrm{PHASE1}_{ith} + \delta_2\mathrm{PHASE2}_{ith}\Big) \times \bar{y_0} + \epsilon_{ith} \qquad (2) \end{aligned}$$

In this specification, we include device, date and hour-of-day fixed effects as in equation (1). $\bar{y_0}$ is the mean water use of each shower head in the baseline period, but normalized by subtracting global mean water use (in both residential colleges) in the baseline period. The main coefficients of interest are the γ terms, which tell us how the treatment effects vary with mean baseline use. To account for differential effects by mean baseline use in both phases, we also interacted PHASE1$_{ith}$ and PHASE2$_{ith}$ dummies with $\bar{y_0}$ respectively. Table 5 presents the results.

For the average baseline user, the reported treatment effects are largely similar in each of the experimental conditions. Notwithstanding, it is worth highlighting that there is a clear divergence in treatment effects between the 18L and 28L conditions in phase 2, when the assigned goal converged to 24L. In particular, we now have sufficient power to reject the null hypothesis of equality at the 1% level ($p = 0.011$), which supports our main finding in Table 3.

Zooming in on the γ terms, we find striking differences between the 18L and 28L conditions. In phase 1, while both conditions display similar conservation effects of between 4.7 and 5.2 liters per shower, the underlying treatment dynamics by baseline use stand in stark contrast to each other. More concretely, we observe that there is only a marginal interaction effect, if any, for the 18L condition ($\hat{\gamma}_{18L1} = -0.105; p = 0.073$) but a highly significant effect for the 28L condition ($\hat{\gamma}_{28L1} = -0.327; p = 0.000$). For the latter condition, a one liter increase in mean baseline use increases the treatment effect by approximately 0.33 liters, which amounts to additional gains of 6.3%. We can easily reject the null hypothesis of equality of the interaction terms ($p = 0.002$). This leads us to conclude that while the average treatment effects in phase 1 are similar, the different goals (18L vs. 28L) had in

Table 5: *Interaction effects with baseline water use*

Dependent variable:	Water use per shower (liters)	
	PHASE 1 (1)	PHASE 2 (2)
MS × PHASE	– 0.776 (0.523)	– 0.870 (0.845)
18L × PHASE	– 4.699*** (0.535)	– 3.885*** (0.877)
28L × PHASE	– 5.170*** (0.605)	– 6.058*** (0.844)
MS × PHASE × $\bar{y_0}$	0.070 (0.050)	0.167* (0.098)
18L × PHASE × $\bar{y_0}$	– 0.105* (0.058)	– 0.107 (0.067)
28L × PHASE × $\bar{y_0}$	– 0.327*** (0.066)	– 0.388*** (0.086)
PHASE × $\bar{y_0}$	– 0.022 (0.032)	– 0.137 (0.085)
Constant	33.908*** (0.244)	
p-value for $\beta_{18L} = \beta_{28L}$	0.455	0.011
p-value for $\beta_{18L1} = \beta_{18L2}$	0.181	
p-value for $\beta_{28L1} = \beta_{28L2}$	0.161	
p-value for $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$	0.003	
p-value for $\gamma_{18L} = \gamma_{28L}$	0.002	0.001
p-value for F-test	0.000	
Device FEs	*Yes*	
Date FEs	*Yes*	
Hour-of-day FEs	*Yes*	
R^2	0.138	
Observations	114627	

Notes. This table shows the interaction effects of moral suasion and real-time feedback (with mean baseline use) on water use per shower. The results are obtained by estimating equation (2) that includes controls for device, day and hour dummies. For ease of interpretation, we split the estimates from a single regression into 2 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. *live* showers) is reported in the last row. Standard errors clustered at the residence × floor × bathroom type level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

fact induced differing adjustments for residents with different baseline water use behavior. Interestingly, we see that the estimates hold steady in phase 2 when the goal converged to 24L for both experimental conditions. While there is no significant interaction effect for the 18L condition ($\hat{\gamma}_{18L2} = -0.107; p = 0.114$), we continue to observe a highly significant and quantitatively large effect for the 28L condition ($\hat{\gamma}_{28L2} = -0.388; p = 0.000$). It is also evident that the interaction effects in phase 2 are significantly different from each other ($p = 0.001$).

This set of differing interaction effects is a novel contribution to the existing literature on resource conservation relating to the provision of real-time feedback. In our experimental context, the real-time feedback intervention has a more muted response by mean baseline use, when coupled with an overly-ambitious goal of 18L. Interestingly, this diminished interaction effect persists even after the goal had been adjusted upwards to 24L in phase 2. We can rule out the hypothesis that the users had simply ignored the change of goal and feedback in phase 2, as we do observe the respective average treatment effects pulling apart for both conditions. Taken together, the evidence strongly suggests that the effectiveness of the intervention hinges on the initial goal assignment, and in the case of an overly-ambitious goal, the subjects may find themselves stuck in a sub-optimal steady state.

4 Robustness Checks

In this section, we conduct a series of tests to bolster the results presented above, and rule out alternative explanations for the observed treatment effects.

4.1 Sample Selection

We begin by addressing the concern about our sample selection, which only includes live showers instead of all recorded showers. We offer three main reasons for focusing our analysis on live shower events only. First, we have the actual date and time when a shower was taken, and this information is necessary for estimating the date and hour-of-day fixed effects precisely. Second, the smart shower head has to be sufficiently powered by water flow to connect to our server and transmit real-time data. By the same token, we can be certain that the feedback lights were working properly in these live shower events since the shower head must be powered on. Finally, live showers constitute up to 91% of all recorded showers, so the restricted sample should remain representative of all shower events.

Notwithstanding, we re-estimate equation (1) using the full sample of recorded showers of strictly more than 4 liters. The results are presented in Table A1, which offers a useful check for whether our sample selection is representative of the full sample. It is evident that the general result continues to hold true. In particular, we still observe similar average treatment effects between the 18L and 28L conditions in phase 1, followed by a divergence in phase 2. The slight difference is that we now have less power to reject the null hypothesis of equal treatment effects between 18L and 28L conditions in phase 2 ($p = 0.060$). This is not surprising as we are using offline shower data which does not contain information about the actual date and time each shower was taken. To estimate the date and hour-of-day fixed effects here, we could only use the date and time of data upload for each shower event, which is an imperfect proxy at best.[14] Therefore, this exercise provides justification for our sample selection of live showers only.

4.2 Fraction of time spent under flashing red light

In the transition from phase 1 to 2, we replaced the shower posters to reflect the new goal of 24L, and remotely configured the new feedback lights from our server. A concern that may arise is that the residents were simply not aware of the change of goal, which potentially undermines the treatment in phase 2. However, the fact that we found a divergence in treatment effects in phase 2 suggests that (at least) a non-trivial proportion of residents were aware and responded to the new goal. If anything, our reported treatment effect gap between the *18L GOAL* and *28L GOAL* groups is a lower bound estimate.

To support the interpretation of diverging treatment effects in phase 2, we consider a related outcome variable, i.e. fraction of time spent under flashing red light. We are able to construct this outcome measure for the sample of live shower events, as we can precisely compute the duration each user spent showering under flashing red light (i.e. beyond the shower goal). Our aim is to show that there are significant changes in the fraction of time spent under flashing red light for the *18L GOAL* and *28L GOAL* groups, from phase 1 to 2. In particular, we want to show that the fraction of time spent under flashing red

[14] The date and time of data upload may not coincide with the actual date and time when each shower was taken. The time lag could span a few hours, and in some cases, up to a few days.

light increased for the *18L GOAL* group in phase 2, which would be consistent with the reduction in treatment effect.

In the succeeding analysis, we only consider live shower observations in phase 1 and 2. In the first exercise, we restrict the sample to observations from the *MS* and *18L GOAL* groups, during the intervention period. For the *MS* group which would henceforth serve as our "control", we define the time spent under flashing red light by assigning "placebo" goals to observations in the respective phases (i.e. 18L in phase 1 and 24L in phase 2). Analogously, for the second exercise, we only take observations from the *MS* and *28L* GOAL groups, and define the "placebo" goals accordingly (i.e. 28L in phase 1 and 24L in phase 2). Formally, we estimate the following fixed effects model which includes interactions with mean baseline use.

$$\begin{aligned} y_{ith} = \alpha_i + \lambda_t + \gamma_h + \beta\Big(\text{TREAT}_i \times \text{PHASE2}_{ith}\Big) + \delta\Big(\text{PHASE2}_{ith} \times \bar{y_0}\Big) \\ + \gamma\Big(\text{TREAT}_i \times \text{PHASE2}_{it} \times \bar{y_0}\Big) + \epsilon_{ith} \end{aligned} \tag{3}$$

In this specification, y_{ith} is the outcome variable, i.e. fraction of time spent under flashing red light. TREAT_i is a dummy variable that equals one for the *18L GOAL* (respectively, *28L GOAL*) group for the former (latter) exercise. The *MS* group serves as our "control" and phase 1 is the omitted time period.[15] As in equations (1) and (2), we include device, date and hour-of-day fixed effects. Therefore, β is the coefficient of interest, which represents the average treatment effect on fraction of time spent under flashing red light for the average baseline user. Table 6 presents the results.

As is evident from column 1, we find a significant increase in fraction of time spent under flashing red light (2.5%) in phase 2, for the average baseline user assigned to the *18L GOAL* group. Conversely, in column 2, we observe a significant reduction (1.3%) for the average baseline user in the *28L GOAL* group. This suggests that subjects in the *18L GOAL* (respectively, *28L GOAL*) group increased (decreased) their shower water use in phase 2, relative to phase 1. This is thus consistent with the interpretation of diverging treatment effects on the intensive margin.

[15] We do not consider shower observations from the baseline period.

Table 6: *Effects of changing goals on time spent under flashing red light*

Dependent variable:	Fraction of time spent under flashing red light	
	18L GOAL (1)	28L GOAL (2)
TREAT × PHASE_2	0.025***	– 0.013**
	(0.008)	(0.007)
$\text{PHASE}_2 \times \bar{y_0}$	– 0.001	0.001
	(0.001)	(0.001)
TREAT × $\text{PHASE}_2 \times \bar{y_0}$	0.001	0.000
	(0.001)	(0.001)
p-value for F-test	0.000	0.004
Device FEs	Yes	Yes
Date FEs	Yes	Yes
Hour-of-day FEs	Yes	Yes
Mean dependent variable (phase 1)	0.302	0.148
R^2	0.168	0.155
Observations	31054	30069

Notes. This table reports the effects of changing goals on fraction of time spent under flashing red light. The results are obtained by estimating equation (3) that includes controls for device, day and hour-of-day dummies. Column 1 reports estimates from the sample of *MS* and *18L GOAL* groups while column 2 reports estimates from the sample of *MS* and *28L GOAL* groups. In both regressions, the *MS* group serves as the "control" and phase 1 is the omitted time period. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. live showers) is reported in the last row. Standard errors clustered at the residence × floor × bathroom type level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.3 Efficacy of stand-alone feedback lights

Another potential explanation for our observed treatment effects (especially in phase 1) is that users would still have responded to the feedback lights, even if no actual information about their water use is conveyed through the shower poster and changing of lights at specified thresholds. This might explain why we observe that the treatment effects are essentially the same for the 18L GOAL and 28L GOAL groups in phase 1. However, we can effectively rule out this argument because in a related experiment, we find that the display of incongruent feedback lights (i.e. no meaningful information about water use) only induced a small conservation effect (i.e. roughly 1/3). This highlights that the large effects observed in our treatments are contingent on the implementation of coherent information feedback.

5 Conclusion

Goals are popular tools for shaping motivation, effort provision and performance, both in the private and public sectors. However, if goals inherit the properties of reference points, changing them may pose challenges and potentially lead to side effects. Moreover, with the growing use of real-time feedback interventions to meet resource conservation targets, it has become even more vital to understand the underlying mechanisms associated with goal setting.

To this end, we conducted a randomized field experiment in two residential colleges at the National University of Singapore to examine the effects of setting goals and changing the level of difficulty over time. We implemented two key treatment groups: the former was assigned a hard goal (18L), while the latter was assigned a moderate goal (28L) in phase 1. Subsequently, both groups received the common, intermediate goal (24L) in phase 2. In other words, both groups started with different goals, but midway through the intervention, had their respective goals adjusted (either upwards or downwards) to the same level of difficulty.

We find that the use of real-time feedback relative to a goal (above and beyond moral suasion) induces large and significant reductions in shower water use. Notably, in phase 1, both the *18L GOAL* and *28L GOAL* groups performed equally well on average. However, in phase 2, when the initial goals are adjusted to the same level of difficulty (24L), differences in performance appear: the *18L GOAL* group now systematically performs worse than the *28L GOAL* group. This suggests that setting too hard a goal might not necessarily lead to immediate damage (at the aggregate level), but might only be manifested in the longer term when goals are changed.

Notwithstanding, the fact that the average treatment effects are the same in phase 1 masks heterogeneity with respect to baseline water use. In particular, we find differential interaction effects with baseline water use between the *18L GOAL* and *28L GOAL* groups from the outset of phase 1. While there is only a marginal interaction effect for the *18L GOAL* group, we find a highly significant interaction effect for the *28L GOAL* group. Strikingly, this heterogeneous effect carries over to phase 2, even though both groups are now assigned the same goal. This suggests that setting an initial hard goal can permanently diminish effort and performance, which is a novel finding that goes beyond our behavioral predictions.

Several mechanisms can generate the observed behavior in our results. It could either be due to loss aversion or fixed penalty around the goals (both give similar predictions), or possibly from psychological disengagement when goals are too difficult. Suppose the effects are due to reference dependence or fixed penalty, the outcome would depend on which direction the goal is adjusted from the initial hard goal. For psychological disengagement, it would not matter which direction the hard goal is adjusted as the subjects just stop paying attention to the goal. We are unable to distinguish between the mechanisms in this paper since we do not further increase the level of difficulty of the initial hard goal (18L). Making a distinction between the underlying behavioral mechanisms is a promising avenue for future research.

Our findings have important implications for policy makers and management since they underscore the importance of selecting an optimal goal. There are two main takeaways from our results: First, in resource conservation, the role of goals may be particularly efficacious in domains where marginal costs to individuals are low (or zero, in our setting), as is often the case for water or energy use under certain rental agreements. In these settings, there might be little (or no) monetary incentives to save water or energy. 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, the provision of real-time feedback relative to a goal can serve as a powerful behavioral tool for resource conservation, even without percuniary motivations. Our work is in line with Tiefenbeck et al. (2019), who show that real-time feedback (on energy use) during showers leads to higher energy-savings per shower among hotel guests, in a similar setting with no monetary incentives involved. This provides policy makers with a new set of non-price interventions to promote resource conservation, which goes beyond the traditional (percuniary) approach of conservation taxes, rebate programs and subsidies.

Second, our results suggest that past goals affect may current effort provision. The goal difficulty should be chosen at an appropriate level, keeping in mind that it would not only affect performance in the current period, but potentially in future periods. Setting too hard a goal might not only lead to suboptimal performance in the near term, but also lead to permanent (negative) effects that cannot be easily undone by simply changing the goal. Further research on the dynamic effect of goal-setting is needed to shed more light on the underlying interactions and mechanisms of our findings.

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.

Allen, E. J., Dechow, P. M., Pope, D. G., and Wu, G. (2017). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63(6):1657–1672.

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.

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.

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

Byrne, D., Goette, L., Martin, L., Schoeb, S., Tiefenbeck, V., and Staake, T. (2018). The behavioral mechanisms of habit formation: evidence from a field experiment. Technical report, University of Melbourne.

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

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.

Doerr, J. (2018). *Measure What Matters.* Portfolio; Illustrated Edition (April 24, 2018).

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.

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).

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

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

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

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

Herweg, F., Müller, D., and Weinschenk, P. (2010). Binary payment schemes: Moral hazard and loss aversion. *American Economic Review*, 100(5):2451–77.

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. and Latham, G. P. (1990). *A theory of goal setting & task performance.* Prentice-Hall, Inc.

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. and Latham, G. P. (2006). New directions in goal-setting theory. *Current directions in psychological science*, 15(5):265–268.

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.

Stigler, G. J. and Becker, G. S. (1977). De Gustibus Non Est Disputandum. *American Economic Review*, 67(2):76–90.

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.

Wu, G., Heath, C., and Larrick, R. (2008). A prospect theory model of goal behavior. *Unpublished manuscript.*

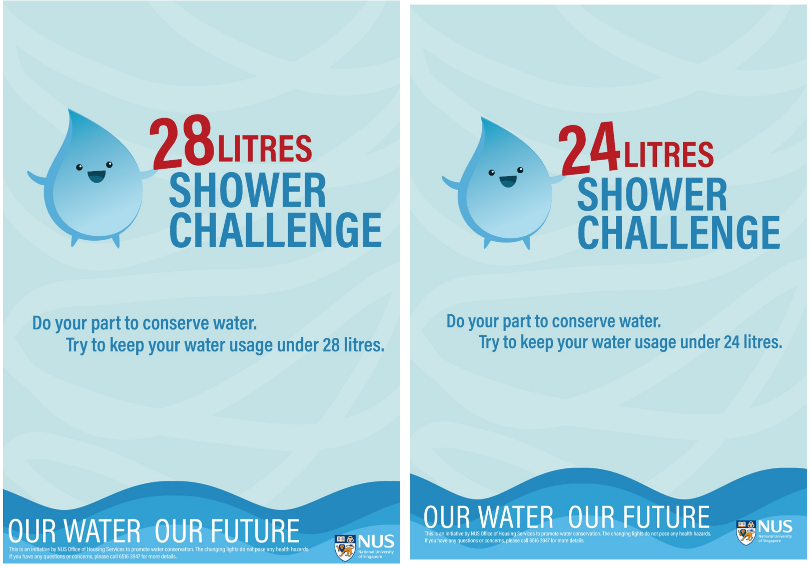
Appendix

(a) *Tembusu and Cinnamon colleges*

(b) *Single room*

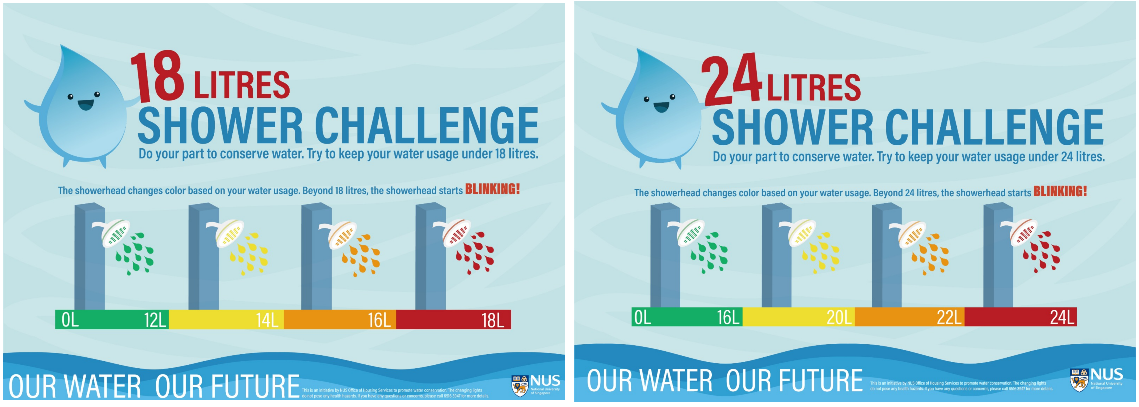
(c) *Bathroom with 2 shower facilties*

Figure A1: *Experimental site*


28 LITRES
SHOWER CHALLENGE
Do your part to conserve water.
Try to keep your water usage under 28 litres.
OUR WATER OUR FUTURE
This is an initiative by NUS Office of Housing Services to promote water conservation. The changing lights do not pose any health hazards.
If you have any questions or concerns, please call 6516 3947 for more details.
NUS National University of Singapore
24 LITRES
SHOWER CHALLENGE
Do your part to conserve water.
Try to keep your water usage under 24 litres.
OUR WATER OUR FUTURE
This is an initiative by NUS Office of Housing Services to promote water conservation. The changing lights do not pose any health hazards.
If you have any questions or concerns, please call 6516 3947 for more details.
NUS National University of Singapore

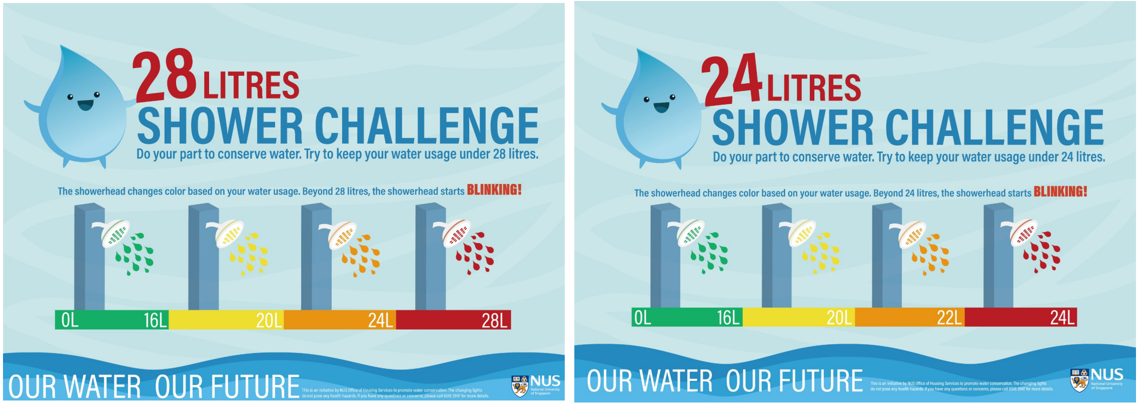
(a) *Phase 1* **(b)** *Phase 2*

Figure A2: *Posters for the Moral Suasion group*


18 LITRES
SHOWER CHALLENGE
Do your part to conserve water. Try to keep your water usage under 18 litres.
The showerhead changes color based on your water usage. Beyond 18 litres, the showerhead starts BLINKING!
0L
12L
14L
16L
18L
OUR WATER OUR FUTURE
NUS
24 LITRES
SHOWER CHALLENGE
Do your part to conserve water. Try to keep your water usage under 24 litres.
The showerhead changes color based on your water usage. Beyond 24 litres, the showerhead starts BLINKING!
0L
16L
20L
22L
24L
OUR WATER OUR FUTURE
NUS

(a) *Phase 1* **(b)** *Phase 2*

Figure A3: *Posters for the 18L GOAL group*


28 LITRES
SHOWER CHALLENGE
Do your part to conserve water. Try to keep your water usage under 28 litres.
The showerhead changes color based on your water usage. Beyond 28 litres, the showerhead starts BLINKING!
0L
16L
20L
24L
28L
OUR WATER OUR FUTURE
NUS
24 LITRES
SHOWER CHALLENGE
Do your part to conserve water. Try to keep your water usage under 24 litres.
The showerhead changes color based on your water usage. Beyond 24 litres, the showerhead starts BLINKING!
0L
16L
20L
22L
24L
OUR WATER OUR FUTURE
NUS

(a) *Phase 1* **(b)** *Phase 2*

Figure A4: *Posters for the 28L GOAL group*

Table A1: *Effect of moral suasion and real-time feedback on water use per shower using full sample of recorded showers*

Dependent variable:	Water use per shower (liters)	
	PHASE 1 (1)	PHASE 2 (2)
MS × PHASE	– 0.702 (0.489)	– 1.297 (0.800)
18L × PHASE	– 4.666*** (0.523)	– 3.590*** (0.812)
28L × PHASE	– 4.619*** (0.630)	– 4.922*** (0.753)
Constant	33.674*** (0.246)	
p-value for $\beta_{18L} = \beta_{28L}$	0.945	0.113
p-value for $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$	0.006	
p-value for F-test	0.000	
Device FEs	*Yes*	
Date FEs (using *offline* date)	*Yes*	
Hour-of-day FEs (using *offline* hour)	*Yes*	
R^2	0.130	
Observations	128323	

Notes. This table shows the effects of moral suasion and real-time feedback (based off different shower goals in each phase) on water use per shower. The results are obtained by re-estimating equation (1) using the full sample of recorded showers. We include controls for device, day and hour-of-day dummies. For ease of interpretation, we split the estimates from a single regression into 2 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. offline + live showers) is reported in the last row. Standard errors clustered at the residence × floor × bathroom type level in parentheses.

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