Fang, Ximeng et al.

Working Paper

Complementarities in behavioral interventions: Evidence from a field experiment on resource conservation

Discussion Papers of the Max Planck Institute for Research on Collective Goods, No. 2023/13

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Fang, Ximeng et al. (2023) : Complementarities in behavioral interventions: Evidence from a field experiment on resource conservation, Discussion Papers of the Max Planck Institute for Research on Collective Goods, No. 2023/13, Max Planck Institute for Research on Collective Goods, Bonn,
https://hdl.handle.net/21.11116/0000-000D-F88E-C

This Version is available at:
https://hdl.handle.net/10419/283136

Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
2023/13

Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Resource Conservation

Ximeng Fang
Lorenz Goette
Bettina Rockenbach
Matthias Sutter
Verena Tiefenbeck
Samuel Schoeb
Thorsten Staake

MAX PLANCK
SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods

2023/13

Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Resource Conservation

Ximeng Fang / Lorenz Goette / Bettina Rockenbach / Matthias Sutter / Verena Tiefenbec / Samuel Schoeb/ Thorsten Staake

November 2023

Forthcoming in *Journal of Public Economics*

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
https://www.coll.mpg.de

Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Resource Conservation*

Ximeng Fang Lorenz Goette Bettina Rockenbach Matthias Sutter
Verena Tiefenbeck Samuel Schoeb Thorsten Staake ‡

30 August 2023

Abstract

Behavioral policy often aims at influencing behavior by mitigating biases due to, e.g., imperfect information or inattention. We study how this is affected by the simultaneous presence of multiple biases arising from different sources, through a field experiment on resource conservation in an energy- and water-intensive everyday activity (showering). One intervention, shower energy reports, primarily targeted knowledge about environmental impacts; another intervention, real-time feedback, primarily targeted salience of resource use. We find a striking complementarity. While only the latter induced significant conservation effects when implemented in isolation, each intervention became more effective when implemented jointly. This is consistent with predictions from a theoretical framework that highlights the importance of targeting all relevant sources of bias to achieve behavioral change.

JEL classification: D83, D90, Q41

Keywords: behavioral public policy, pro-environmental behavior, limited attention, information provision, real-time feedback, policy interactions

*We are indebted to our research assistant team (in particular Benedikt Kauf and Kristina Steinbrecher) for their indefatigable support in running the field experiment, and to Lim Zhi Hao for editing. We would further like to thank audiences in Bonn, at the EEA Virtual 2020, the ECONtribute Retreat 2020, the ESA Dijon, the IAREP-SABE Dublin, the EAERE Manchester, the 3rd CRC TR 224 Conference, the PCBS 2019 Prague, the 12th RGS Doctoral Conference, and the 2nd BoMa PhD Workshop for helpful comments. Financial support by the University of Cologne Forum "Energy", the SUN Institute Environment & Sustainability, and by the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B07) is gratefully acknowledged. The 2019 supplementary survey was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanyt's Excellence Strategy EXC 2126/1 390838866. This study is registered in the AEA RCT Registry under the identification number "AEARCTR-0004865".

‡Fang: University of Oxford (email: ximeng.fang@sbs.ox.ac.uk). Goette: University of Bonn, National University of Singapore. Rockenbach: University of Cologne. Sutter: Max Planck Institute for Research into Collective Goods, University of Cologne, University of Innsbruck. Tiefenbeck: Friedrich-Alexander Universität Nürnberg-Erlangen, ETH Zurich. Schoeb: University of Bamberg. Staake: University of Bamberg, ETH Zurich.

1. Introduction

Amidst growing concern about climate change and resource scarcity, many individuals intend to make personal sacrifices to protect the environment; yet they often fail to act pro-environmentally in their everyday lives (Kollmuss and Agyeman, 2002; Frederiks et al., 2015). This gap between intentions and actions can result from a multiplicity of behavioral frictions and biases. For instance, consumers tend to underestimate the impact of highly resource-intensive activities (Attari et al., 2010; Attari, 2014; Imai et al., 2022), and they may also not be fully attentive to their resource use (Allcott, 2016; Tiefenbeck et al., 2018). Importantly, when biased behavior arises from multiple different sources at the same time, this could not only prevent individuals from acting on their intrinsic prosocial or pro-environmental motives, but also mute their response to policy interventions that only address a subset of all relevant biases.

This problem that comes with multiple biases is reminiscent of the Anna Karenina principle, which states that failure in just one factor out of many can lead to failure of an objective as a whole.[1] For example, providing information to correct misperceptions of environmental impacts has little effect on behavior if individuals remain inattentive and exhibit self-control problems, status quo bias, and so on.

Conversely, drawing attention to environmental impacts only has a muted effect on behavior if agents remain unaware of the true extent of the externalities caused by their actions. In this example, addressing both information problems and biases due to, e.g., limited attention, could produce synergies in the form of positive interaction effects. More generally, combining interventions that focus on different biases each could result in complementarities, defined as *each* intervention becoming more effective when implemented in conjunction with the other(s) than in isolation (Coe and Snower, 1997). While the use of combined interventions is widespread in (behavioral) public policy, less is known about when and why one may expect interventions to be complements, which can be crucial for guiding effective policy design. In this paper, we highlight the role of multiple biases.

We report evidence from a three-month randomized field experiment in which we used two well-studied behavioral policy tools to encourage resource conservation in an energy- and water-intensive everyday activity, namely showering. Our interventions were designed in such a way that they target different potential sources of biased behavior. The first intervention, shower energy reports, inspired by the Opower home energy reports (Allcott, 2011), were primarily aimed at closing knowledge gaps about environmental impacts by providing information on water use as well as on energy use and CO_2 emissions due to water heating. The second intervention, real-time feedback, provided immediately visible and salient information on water consumption — but not energy use

[1] The Anna Karenina principle is inspired by the opening phrase of Leo Tolstoy's novel *Anna Karenina*: "All happy families are alike; each unhappy family is unhappy in its own way." (Tolstoy, 2003). One might cheekily adapt this principle to our context by stating a slightly modified version: All unbiased agents are alike; all biased agents are biased in their own way.

or CO_2 emissions — through a smart meter display (Tiefenbeck et al., 2018), and could thus help individuals focus their attention while they engaged in the activity. Crucially, we implemented a complete 2×2-design to evaluate both the combined intervention as well as each intervention in isolation. Our main finding is that implementing the interventions jointly seemed to result in a super-additive boost of resource conservation effects compared to their effects when implemented in isolation. This is in line with the idea that both information- and attention-based mechanisms might have been necessary to achieve behavioral change in our context.[2]

To formalize our arguments, we introduce a novel theoretical framework in which overconsumption can arise from multiple sources of bias (e.g., imperfect information, limited attention). Each of the biases acts akin to a discount factor and agents are prevented from incorporating the full marginal costs of resource use by the product of all biases. A key implication is that when agents suffer not just from one but from multiple independent biases (à la Anna Karenina), then interventions can become complements if they each focus on a different behavioral mechanism. The intuition is simple: the more unbiased an agent is in one dimension, the larger is the impact of reducing bias in another dimension. For example, the more attention an agent pays to her resource use behavior, the more likely it is that she will actually change her behavior when learning that the environmental impact is more negative than previously thought. Thus, in this example, mitigating both attention and information problems can have mutually reinforcing effects. This interaction mechanism is absent when two interventions operate mostly through the same behavioral channel, e.g., if they provide the same type of information.

Resource usage in the shower offers a useful context for studying complementarities in behavioral interventions, for several reasons. First, showering is resource-intensive: an average shower in our sample required 2.2 kWh of energy to heat up 38 liters of water, which corresponds to about 10% of the average residential energy use and 30% of the average water consumption per capita and day in Germany, where we conducted our study.[3] Second, most individuals underestimate the CO_2 emissions caused by water heating for showering — by as much as 89% on average based on our own survey data —, which creates scope for conservation through belief correction (Byrne et al., 2018). Third, showering is also prone to behavioral biases like limited attention and self-control problems, as the pleasure of a warm shower is salient and immediate, whereas the cost of resource use seems abstract and is hard to keep track of. Hence, individuals may not

[2]Complementarity can also arise if our interventions do not exactly work through the described mechanisms, as long as they sufficiently differ from each other in their targeted bias. For example, real-time feedback could be interpreted as facilitating learning or optimization, and this information can be complementary to the information on CO_2 emissions provided through shower energy reports.

[3]Source: German Federal Statistical Office (`https://www.destatis.de/EN/Themes/Society-Environment/Environment/Environmental-Economic-Accounting/private-households/Tables/energy-consumption-households.html` and `https://www.destatis.de/EN/Themes/Society-Environment/Environment/Water-Management/_node.html`). About 70% of energy for room and water heating in Germany was generated from fossil fuels at the time of our study (dena, 2016).

fully engage in conservation efforts unless they are informed about the actual impact of their behavior *and* keep environmental concerns on top of their minds while showering.

We conducted our field experiment in student dormitories in the cities of Bonn and Cologne, Germany, in the winter term 2016/17. A total of 351 students participated in our experiment, all of them living in single-person dorm apartments with a private bathroom. For the duration of our study, from early December 2016 until early March 2017, each participant was equipped with a smart shower meter that recorded detailed data of each shower taken. Subjects were randomly assigned into one of four experimental conditions: no intervention (CON group), shower energy reports only (SER group), real-time feedback only (RTF group), or both interventions combined (DUAL group). After an initial baseline stage, the smart meter started displaying real-time feedback on water use for subjects in RTF and DUAL, and about halfway into the study, we further started sending individualized shower energy reports via email to subjects in SER and DUAL, using data uploaded from the smart meters. This staggered design allows us to identify treatment effects of each intervention regime in a difference-in-differences setup.

Our empirical results show that, compared to the control group, subjects in the RTF group reduced their energy (water) consumption by about 0.4 kWh (6.3 liters) per shower, which corresponds to 17–18% of baseline resource use. This treatment effect remains stable over the entire 3-month duration of the study. Energy reports in isolation (SER group) did not lead to any statistically detectable conservation effect. However, in line with our hypothesis, we observe a striking complementarity between the two interventions. Combining energy reports with real-time feedback (DUAL group) *further* increased the treatment effect of real-time feedback in isolation by an additional 0.23 kWh of energy (3.8 liters of water) per shower. Thus, it seems that the shower energy reports in our context could only start to unfold their potential when subjects were in an enhanced choice environment where their resource usage was immediately visible. We find no evidence of adjustments on the extensive margin, i.e., the number of showers people take.

The additional reduction of resource use in the DUAL group was not driven by short-lived boosts directly after receiving a shower energy report, but rather seemed to unfold over time, which speaks against Hawthorne or pure reminder effects as the underlying mechanism. Data from baseline and endline questionnaires shows that both interventions helped subjects form more precise beliefs about their own water use in the shower and that there is no evidence that subjects in the DUAL group read their reports more carefully than subjects in the SER group. Supplementary survey results from a comparable sample further suggest that information included in shower energy reports also induces drastic (upward) updates in beliefs about CO_2 emissions due to warm water consumption in the shower. Hence, the null result for shower energy reports in isolation was unlikely due to lack of learning. Instead, it seems that in the absence of real-time feedback, inattention and lack of immediate visibility have prevented knowledge gains about environmental impacts from translating into effective conservation behavior.

Overall, our findings are consistent with the theoretical argument that in the presence of multiple biases, different behavioral interventions can become complements, because targeting one source of bias (e.g., imperfect information) becomes more effective when residual biases (e.g., inattention, present bias) are also mitigated, and vice versa. One implication is that lack of evidence for effectiveness of an intervention in isolation such as in the case of shower energy reports in our study does not mean that it cannot be effective in an enhanced policy environment that also takes into consideration further behavioral mechanisms. Appropriate policy bundling may thus increase the cost-effectiveness of interventions beyond what can be achieved with piecemeal approaches.

Our study builds on important previous contributions that have investigated the effects of similar interventions on household resource conservation.[4] For example, in an influential evaluation of the Opower home energy reports, which provide information on aggregate electricity use to millions of U.S. households, Allcott (2011) reports an average household-level conservation effect of 2%, or about 0.62 kWh per day, although effects might be smaller in countries with lower baseline consumption (Andor et al., 2020) or when monetary incentives to save energy are low (Myers and Souza, 2019). Our SER intervention is inspired by the these home energy reports. One key difference is that our reports only provide information on one specific activity (showering) instead of aggregate household consumption, as disaggregated feedback could enable better learning and thus stronger conservation responses in the targeted activities (Gerster et al., 2020), in particular when provided in shorter time intervals or even in real time. Tiefenbeck et al. (2018) provide real-time feedback in the shower through the same type of smart meter that we use in this study and document a conservation effect of 22% (0.6 kWh less energy and 9 liters less water per shower). These results also replicate in a sample without monetary incentives and without self-selection into the study (Tiefenbeck et al., 2019). One important mechanism of real-time feedback is that it draws immediate attention to resource consumption by making it more salient (Bordalo et al., 2022). This study addresses the question is whether this can be used to complement interventions that aim to encourage pro-environmental action through other mechanisms, like more detailed information provision or social norms, and could thus benefit from generally higher attention to relevant behaviors.

We further relate to a number of other studies that test a combination of interventions, and especially to studies on pro-environmental behavior that also consider the idea that policy measures might become more effective when implemented in conjunction with others — although it should be noted that some studies lack a complete experimental

[4]Pro-environmental interventions have drawn from a broad set of instruments such as information provision, social norms, goal-setting, etc. For reviews, see e.g. Abrahamse et al. (2005), Fischer (2008), Delmas et al. (2013), Karlin et al. (2015), Andor and Fels (2018), Carlsson et al. (2021), and Khanna et al. (2021). Information provision in particular is often regarded as a promising policy lever, as individuals often misperceive the environmental impact of everyday activities (Attari et al., 2010; Attari, 2014; Camilleri et al., 2019) and tend to engage in relatively ineffective conservation measures (Gardner and Stern, 2008; Tonke, 2019).

design required to identify interaction effects.[5] For example, Jessoe and Rapson (2014) find that peak pricing schemes may only reduce peak electricity usage for households who have been outfitted with in-home-displays. Other recent studies who investigate the combination of financial incentives and behavioral interventions tend to find that they affect behavior along different margins or for different subpopulations, but find no conclusive patterns with regard to interaction effects (List et al., 2017; Holladay et al., 2019; Giaccherini et al., 2020; Fanghella et al., 2021). Hahn et al. (2016) test the individual and combined effects of social comparisons and loss framing on take-up of water-efficient technology as well as general household water consumption, but the results for interaction effects are mixed. Brandon et al. (2019) evaluate the interaction effect of two behavioral interventions on household energy conservation, home energy reports and "peak energy reports", which provide feedback and social norms for households' peak electricity use. As both interventions are very similar and likely operate through similar behavioral channels, it is not clear whether one should expect any interaction effect. Indeed, Brandon et al. find neither strong evidence for complementarity nor substitutability.[6] While we provide a novel case study on the interaction of two specific types of behavioral interventions, our main contribution to this strand of literature is that we attempt to make a step towards understanding mechanisms that systematically lead different policy interventions to become complements, both theoretically and empirically. Specifically, our study highlights the role of multiple biases arising from different sources leading to an Anna Karenina effect. These insights can be adapted to guide hypothesis formation about policy interactions in other contexts as well.

The remainder of this paper is structured as follows: Section 2 introduces the theoretical framework for policy interactions under multiple biases. Section 3 describes the experimental setup and derives behavioral predictions. Section 4 presents our data as well as descriptive statistics. Section 5 lays out our empirical approach and Section 6 presents our main empirical results. In Section 7, we analyze the potential mechanisms underlying the results. Section 8 concludes.

[5] See, for example, the review by Khanna et al. (2021). Combined interventions are also used in other contexts than pro-environmental behavior. For example, in development economics, a number of studies experimentally test the combined effect of different interventions on financial savings (Dupas and Robinson, 2013; Jamison et al., 2014), education (Mbiti et al., 2019), risky sexual behavior (Duflo et al., 2015; Dupas et al., 2018), demand for health products (Ashraf et al., 2013), or immunization (Banerjee et al., 2021). Many of these studies, however, cannot explicitly test policy interactions, as they lack a complete factorial design (Muralidharan et al., 2020), and none of them asks more generally if or why different interventions can be complements if they target separate mechanisms. One notable study is by Mbiti et al. (2019), who find complementarities between providing school grants and adding teacher incentives in improving children's educational outcomes. Another study by Banerjee et al. (2021) employs reminders, incentives, and information ambassador interventions on a large-scale, and then uses a data-driven approach to identify the best combination; in particular, one observation is that information ambassadors seem to amplify the effect of other interventions.

[6] One speculative interpretation is that, if the HERs reduced energy consumption through investments into energy-efficient technology, while the peak energy report reduce energy usage by increasing salience of peak load events, then the potential for complementarity is limited.

2. Theoretical Framework

We begin by introducing a stylized framework to formalize our argument of how complementarities in behavioral interventions can arise in settings where biased behavior arises from multiple sources, e.g. imperfect information, limited attention, present bias.

2.1. Basic Setup

The agent (she) engages in an resource-intensive activity, say showering, and the policy objective is to reduce resource use. Her consumption level is determined by a trade-off between the consumption utility (e.g., hygiene, pleasure, opportunity costs of time) and the perceived costs of resource use (e.g., monetary costs, environmental concern). In the case of showering, both water and energy matter, and each is subject to distinct costs and externalities. For national parsimony, we focus on energy, as it captures water use as well as water heating.[7] Thus, the agent chooses energy use level $e \geq 0$ to maximize

$$U(e) = V(e) - B \cdot ce\,, \tag{1}$$

where $V(e)$ is the instantaneous consumption utility and $c > 0$ is the (constant) marginal cost of energy consumption. We consider a more general convex cost function $C(e)$ in Appendix G. In addition to standard smoothness conditions, we assume that V is hump-shaped (locally increasing at 0, strictly concave, unique maximum). For simplicity, we abstract from uncertainty or dynamics. In the absence of monetary motives, as in our empirical setting, c is the "moral" cost the agent perceives in face of the negative externalities from energy use. However, the perceived cost is attenuated by an aggregate bias factor $B \in [0, 1)$.

Multiple sources of bias. — The aggregate B factor can be the product of a collection of separate factors. Although this is easy to generalize, it is sufficient to focus on a simple case with two sources of bias to illustrate the mechanics:

$$B = b_1 \cdot b_2\,. \tag{2}$$

For example, the first factor b_1 may indicate the degree to which the agent underestimates energy intensity (as shown, e.g., in Attari et al., 2010), and the second factor b_2 the degree to which she is inattentive (e.g., Tiefenbeck et al., 2018). The multiplicative form captures that any single factor can independently prevent the agent from implementing her conservation motive, akin to the Anna Karenina principle. In this example, the agent will not take into account environmental cost both if she believes her behavior has no impact ($b_1 = 0$) *and* if she is fully inattentive ($b_2 = 0$), either condition by itself is sufficient. In the general case of K biases, this would become $B = \prod_{k=1}^{K} b_k$ (see Appendix G.3). Also

[7] Energy for water heating is determined by the amount of water and the temperature gradient. As we will report in section 6, we find no evidence for adjustments in water temperature in our study.

note that, in principle, individuals can be biased towards less consumption (i.e., $b_k > 1$), for example if they overestimate costs (e.g., Wichman, 2017; dAdda et al., 2020).

Consumption behavior. — The agent's choice is defined by the intersection of marginal utility and marginal costs, with the latter being diminished by the aggregate bias:

$$V'(e) \;=\; B \cdot c\,. \tag{3}$$

If $B < 1$, then the marginal cost is underweighted and energy use is thus biased upwards. Correcting the bias (raising B towards 1) thus leads the individual to perceive the cost of consumption more fully. Thus $\frac{\partial e}{\partial B} < 0$, since $V''(e) < 0$.

Behavioral interventions. — In this setup, we define behavioral interventions as policies that aim to change consumers' behavior by changing B. In contrast, price-based policies such as Pigouvian taxes would be aimed at increasing the marginal costs c that the agent faces. As $B = b_1 \cdot b_2$, there are two behavioral policy levers for reducing energy consumption: raising b_1 (e.g. providing information) and raising b_2 (e.g. enhancing salience).

2.2. Policy interaction effects

In our context, two interventions X and Y are complements if their combination reduces behavior by more than the sum of their individual effects: $\Delta e^{XY} \leq \Delta e^{X} + \Delta e^{Y}$. If they are substitutes, the inequality is reversed. Notice that even under substitutability, it can be the case that XY is more effective than either X or Y in isolation, i.e., $\Delta e^{XY} < \Delta e^{X}$ and $\Delta e^{XY} < \Delta e^{Y}$. Thus, to empirically identify interaction effects between different policy interventions, it is necessary to evaluate the effectiveness of each intervention in isolation.

The key mechanism we aim to highlight in this paper is that in the presence of multiple biases, policies that target only one bias dimension may have a limited effect on behavior, whereas the effect of combining several policy levers may be superadditive. For example, correcting perceptions of the environmental impact b_1 may only have a small impact on behavior if the attention parameter b_2 is still close to zero. There is a simple geometric interpretation to illustrate this: the overall bias parameter B, defined in equation (2), can be thought of as the area of a rectangle with sides of lengths b_1 and b_2 (see Figure 1a). The larger the rectangle the lower the resulting energy consumption will be. Now suppose that b_1 is exogenously increased by δ_1. The resulting increase in B will be $\delta_1 b_2$, as it is attenuated by b_2. Analogously, an exogenous increase of δ_2 in the dimension of b_2 results in an aggregate change of $\delta_2 b_1$. The effect of jointly increasing b_1 and b_2 by the same amounts, however, results in an overall change of

$$\Delta B \;=\; \delta_1 b_2 \;+\; \delta_2 b_1 \;+\; \delta_1 \delta_2\,. \tag{4}$$

There is an additional effect of size $\delta_1 \delta_2$, because a gain in one dimension also makes the improvement in the other dimension larger. Geometrically, this is represented by the top

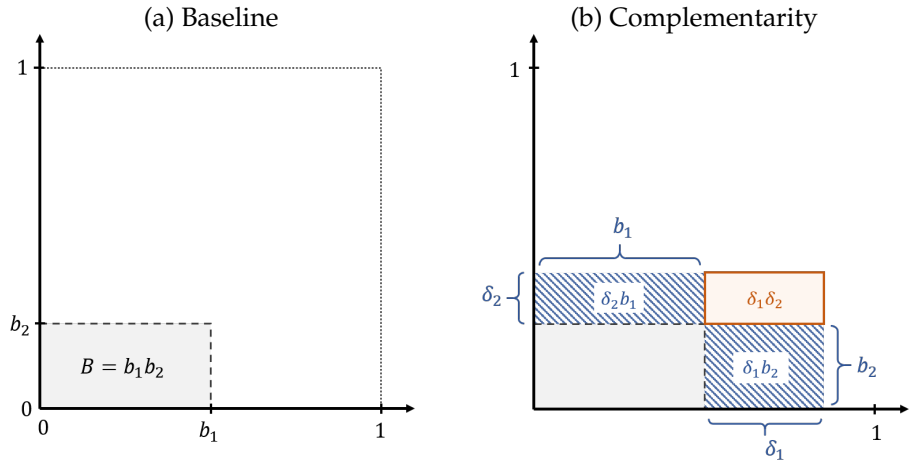


Figure 1: Depiction of example interventions

Notes. The grey rectangle in Figure (a) illustrates the aggregate bias B as defined in equation 2 without any intervention in place. Figure (b) illustrates the increase in B through exogenous interventions in each dimension.

right rectangle outlined in Figure 1b. This mechanism potentially induces complementarity between interventions that mitigate different biases each.[8]

Under what conditions does this complementarity in bias reduction $\delta_1\delta_2$ translate into a complementarity in behavior between interventions X and Y? As formally derived in Appendix G.1, a second-order Taylor approximation yields

$$\Phi^{XY} := \Delta e^{XY} - \Delta e^{X} - \Delta e^{Y} \approx \left[\frac{\partial e}{\partial B} + \frac{\partial^2 e}{\partial B^2} b_1 b_2\right] \delta_1 \delta_2 \,. \tag{5}$$

The term $\frac{\partial e}{\partial B}$ in equation (5) is negative and scales with $\delta_1\delta_2$, thus creating scope for complementarity in behavior. The second term in brackets reflects the change in the slope of $\frac{\partial e}{\partial B}$. Intuitively, one would expect a diminishing responsiveness to bias mitigation ($\frac{\partial^2 e}{\partial B^2} > 0$), as the more the agent already reduces her consumption. This corresponds to $V(e)$ having a positive third derivative. However, if either b_1 or b_2 is sufficiently close to zero, the first-order effect dominates and complementarities in bias reduction translate into complementarities in observable behavior. One might call this the Anna Karenina condition: the more biased an agent is, in multiple ways, the more effective it is to target the biases simultaneously.

2.3. Policy implications

Lastly, we explore implications for policy makers who attach a social cost $\gamma > 0$ to every unit of e (e.g., due to externalities), in addition to the private cost c to the consumer. If the only policy goal was to reduce resource use e, then complementarities in behavior would

[8] We focus here on the case of two "pure" interventions that only target b_1 or b_2, respectively. In practice, many interventions may affect not just one but several biases. In Appendix G, we show that imperfectly targeted interventions can still produce complementarities if sufficiently different from another.

directly carry over to policy benefits. The prevailing view is to define welfare over true consumer utility (e.g., Bernheim and Taubinsky, 2018), so

$$W(e) = V(e) - ce - \gamma e\,. \tag{6}$$

It is straightforward to show that welfare increases as B goes towards 1, as $\frac{\partial W}{\partial B} > 0$ as long as $B < 1$. These welfare gains result from, both, a reduction in externalities and "internalities". Taking a second-order Taylor approximation, we can examine how welfare is affected by combining interventions X and Y:

$$\Delta W^{XY} - \Delta W^{X} - \Delta W^{Y} \;\approx\; \left[(B-1)c - \gamma\right]\Phi^{XY} \;+\; Bc\delta_1\delta_2\,\frac{\partial e}{\partial B} \;\gtreqless\; 0\,. \tag{7}$$

Appendix G.2 contains a detailed derivation. The first term on the right-hand side is positive if $\Phi^{XY} < 0$. However, the second term is always negative, reflecting that gains in consumer surplus decrease as B increases. This implies that complementarity in behavior (i.e., if $\Phi^{XY} < 0$) is a necessary but not sufficient condition for complementarity in welfare. If, in addition, either $\Delta e^{X} \approx 0$ or $\Delta e^{Y} \approx 0$, our model implies that one of the b's is equal to zero. Thus, the joint finding of policy complementarities in behavior *and* one of the interventions being completely ineffective on its own may serve as a sufficient condition that complementarities also exist in terms of welfare.

Equation (7) further suggests that policy makers with a binding budget constraints face a trade-off between targeting a larger share of the population or enriching the policy bundle. For example, suppose the policy maker has a budget g, and that implementing either X or Y in the whole population requires social costs κ^X or κ^Y greater than her budget. Hence, she could cover g/κ^X of households with X, or g/κ^Y of households with Y. Now assume that $\frac{\Delta W^X}{\kappa^X} \geq \frac{\Delta W^Y}{\kappa^Y}$ (X generates a larger welfare gain than Y) and $\frac{\Delta W^X}{\kappa^X} > 1$ (the welfare gain for X is positive). In this case, intervention Y seems unattractive at first glance relative to X. However, our model implies that combining both policies in a bundle (X, Y) at cost $\kappa_X + \kappa_Y$ to fewer households can be preferable to treating a larger number with X in isolation Formally, the condition is $\frac{\Delta W^{XY} - \Delta W^X}{\kappa^Y} \geq \frac{\Delta W^X}{\kappa^X}$. This last condition can only hold if $\Delta W^{XY} - \Delta W^X$ is substantially larger than ΔW^Y, i.e. if complementarities are sufficiently strong.[9] Thus, complementarities can create a unique rationale for unequal policy distribution to improve cost-effectiveness.

[9]One implicit assumption we make here is that the costs of interventions are additive, i.e., $\kappa^{XY} = \kappa^X + \kappa^Y$. Costs can in principle be super- or sub-additive, purely for technological reasons, and hence we abstract from this here. For example, in our study, costs of recruiting participants and setting up the technical infrastructure are fixed, leading to economies of scope when adding interventions.

3. Experimental setup

Our field experiment was conducted from early December 2016 to late February/early March 2017 in a sample of students living in dormitory apartments. Each participant was equipped with a smart meter that measured individual energy and water consumption in the shower over the entire study duration. We then evaluated the effect of two different interventions, real-time feedback and shower energy reports, on resource conservation behavior. To test for complementarity, we further implemented a combined intervention in which subjects received both real-time feedback and shower energy reports.

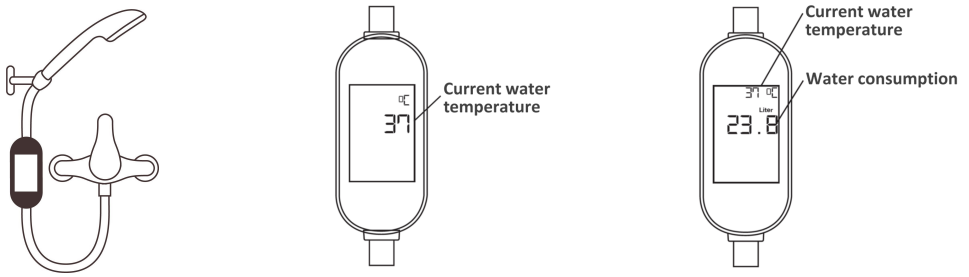
3.1. Recruitment of participants

We selected six student dormitory sites in Bonn and Cologne for our sample, and ran the study from early December 2016 to early March 2017. All dormitory residents were students at the University of Bonn, the University of Cologne, or at various smaller universities in the cities. We recruited our subjects from the pool of dorm tenants living in single-person apartments with private bathroom, as this allows us to precisely measure the resource use of each individual. One noteworthy feature of our sample is that subjects have no direct monetary incentives to conserve energy or water, because they pay a flat monthly rent that includes all utility bills. Hence, any observed conservation response would be solely driven by non-monetary motives and unconfounded by income effects.

To participate in the study, residents had to actively agree based on the principle of informed consent. Two additional criteria were levied: subject should not have lengthy absences planned within the intended study period (except during Christmas vacation), and they should own a smartphone compatible with Bluetooth 4.0, which was necessary for implementing the shower energy reports.

The recruiting process started around mid-October 2016. Posters and flyers informed residents of the selected dormitories about the upcoming study, and our local research assistant teams engaged in door-to-door recruiting. Interested students had to complete an online registration survey to provide required information and to give their consent to the collection and analysis of data on their showering behavior. It was explicitly (and truthfully) stated that we would treat any collected data confidentially and not share it with the dormitory administration. As remuneration, each participant received 20 Euros after completing the study, and ten participants were randomly drawn to receive a 300 Euro cash prize. In total, 406 students registered for the study, out of which 361 met our participation criteria.[10] Ten students subsequently dropped out of the study, either because they moved out of their dorm unexpectedly or because we were not able to contact them again. This leaves us with a final sample of 351 participants.

[10]The total number of all single apartments in the selected dorms was 1380 (vacancies included), thus our gross recruitment rate was about 30%. For more than half of these apartments, we never encountered the resident, so out of the students we actually managed to talk to, the majority registered for the study.



(a) Position in the shower (b) Display in control mode (c) Display in feedback mode

Figure 3: Amphiro b1 smart shower meter

3.2. Smart shower meters and smartphone app

At the beginning of the study, starting on 5th Dec 2016, each participant was equipped with an Amphiro b1 smart shower meter that measures and records data of every water extraction in the shower. The device can be easily attached below the shower head and features a smartphone-sized liquid crystal display, which can be programmed to show various types of information (see Figure 3a). The smart meter is small, lightweight, and needs no battery; power is generated through an integrated hydro turbine, without noticeably affecting water flow in the process. One drawback of the lack of battery is that the device is unaware of the absolute time of day: showers can only be recorded in temporal order, but without time stamps. Once the water flow in the shower starts, the smart meter is powered and begins to measure, among others, the amount of water flowing through, water temperature, and the time passed since beginning of water flow. When water flow stops, the device and its display initially remain switched on, and if water is turned on again within three minutes, it will continue its measurement seamlessly. This accounts for short breaks in water flow when applying soap or shampoo. Once water flow stops for more than three minutes, the device terminates measurement and stores the recorded information to a new data point.

We programmed the shower meters to display select pieces of information to participants in real-time, i.e., while they are taking their showers, contingent on the study progress and assigned experimental condition (as described below). In addition, we asked all participants to install the Amphiro smartphone app around week 5 of the experiment, shortly after the end of the Christmas break. The participants could use the app to upload data from their shower meters via Bluetooth.[11] We were then able to access the uploaded data and use it to create personalized shower energy reports. The original Amphiro smartphone app also calculates summary statistics about users' resource use in the shower, but we deactivated this feature for our study participants, so its only functionality was data uploading. One ancillary benefit of the app was that it stored time and

[11]The process was quite simple. After installing the smartphone app, subjects created an account and paired it to their shower meter. After successful pairing, the meter automatically transmitted all stored data to the app via Bluetooth whenever it was powered on and the smartphone within range.

date of each data upload, which allows us to construct approximate time windows for each shower. About three out of four participants (72%) uploaded all data successfully, while the remaining experienced some technical problems. The most common sources of failure were problems with the Bluetooth connection or unexpected incompatibility between smartphone and app. We will come to back to this issue again later.

3.3. Implementation of real-time feedback

The live tracking of water use on the shower meter display in feedback mode is what we refer to as real-time feedback, our first type of intervention. We programmed half of the smart meters as control devices and the other half as treatment devices. Control devices only displayed the current water temperature throughout the entire study (Figure 3b). Treatment devices also started in control mode for the first ten showers, which we use to measure baseline behavior, but switched permanently to feedback mode starting from the eleventh shower. In feedback mode, the display shows both the water temperature and the amount of water used (in liters) at any time of the shower (Figure 3c). Note that the smart meters did not provide any information on the impact of water temperature on energy consumption or on the energy-intensity of water heating more generally.

3.4. Implementation of shower energy reports

Our second intervention consists of two personalized shower energy reports. These reports were sent via e-mail and showed descriptive statistics about the subject's water and energy use in the shower, as well as information about environmental impacts. Temperature information was not included, as all subjects received this through their smart meter anyway. To allow for learning about outcomes of single showers, a graphical representation of the subject's history of water use per shower was included. The reports were constructed based on data that was uploaded by subjects through the smartphone app. We sent out additional reminders to upload data before each planned delivery, but the reports themselves were not explicitly announced. Subjects who did not manage to upload any data received a report template with blanks in place of statistical figures and graphs.

Appendix Figure A1 shows the screenshot of a typical shower energy report. After a short introductory text, subjects see a scatter plot of their history of water use per shower since the beginning of the study, including a fitted regression line to help recognize trends and averages. Below the graph, average water use (in liters) and energy use (in kWh) per shower are stated numerically. Furthermore, there is a paragraph with information on projected CO_2 emissions per year and the number of trees required to absorb the corresponding amount of CO_2. The report is formulated concisely in neutral language, to avoid any normative or moral suasion elements. In the second report, we added a social comparison component in the spirit of Allcott (2011) and Ferraro and Price (2013), see Appendix Figure A2. Specifically, we assigned a random anonymous peer to each

subject and displayed statistics on the peer's energy and water use.[12] At the bottom of each report, we included personalized link to a mini-survey that we asked subjects to fill out. The mini-survey contained three questions to elicit subjects' estimate of their water consumption per shower (absolute and relative to others). The purpose of this was twofold. First, we use the responses to verify if a subject has read the email carefully and, based on the estimate accuracy, how closely he or she paid attention to the information. Second, we use the time of survey response to determine when exactly a subject read the email.

3.5. Experimental design

We implemented a complete 2×2 design with four experimental conditions. Subjects in the control (CON) group received no intervention at all; subjects in the RTF group only received real-time feedback through the smart shower meters; subjects in the SER group only received shower energy reports; and subjects in the DUAL group received both real-time feedback and shower energy reports. Treatment assignment was randomized and the group sizes are as follows: 82 in CON, 88 in SER, 90 in RTF, 91 in DUAL.[13]

Figure 4 illustrates the experimental design in detail. Each shower meter went through a baseline stage of ten showers, in which it only displayed the current water temperature, regardless of the experimental condition. We use these showers to measure baseline consumption. Starting from the eleventh shower (intervention stage), devices in RTF and DUAL additionally displayed water use in real-time, whereas devices in CON and SER stayed in control mode. About halfway into the study, we started sending energy reports to each subject in the SER or DUAL group; the first report was sent on 24 January 2017 and the second report on 8 February 2017, about two weeks later. We distinguish between intervention (IN) stage 1, in which real-time feedback is switched on but there were no reports yet, and intervention (IN) stage 2, which is the period that begins after subjects saw the first report.[14] In order to hold interaction with experimenters constant, subjects in CON and RTF groups received placebo emails at the exact same time the shower energy reports to subjects in SER and DUAL were sent out. These subjects were asked to complete the same mini-survey that came along with the actual reports.

This staggered experimental design allows us to exploit both between- and within-subject variation to cleanly identify and efficiently estimate treatment effects of interest. The effect of real-time feedback in isolation is identified by the comparison between the RTF and CON groups in the (entire) intervention stage, or alternatively by the comparison between the pooled RTF/DUAL group and the pooled CON/SER group in IN stage

[12] The matching procedure was one-sided and ensured that each subject (except the most and the least efficient) was equally likely to see a peer with lower or higher energy use per shower.

[13] For the exact randomization protocol, see Appendix B.

[14] In practice, the distinction between IN stage 1 and 2 is not perfect, as we observe 23 subjects in our sample who had yet to complete all 10 baseline showers when the first report was sent out. If anything, this generates measurement error in our treatment indicators and thus biases estimates toward zero.

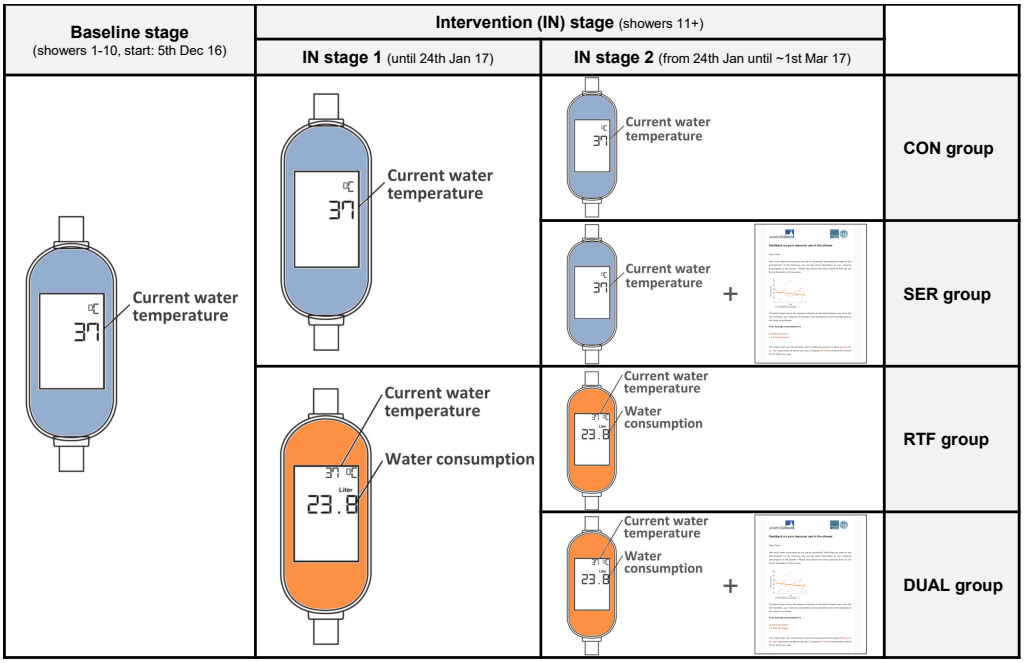


Figure 4: Experimental design and timing of interventions

1. The effect of shower energy reports in isolation is identified by the comparison between the SER and CON groups in IN stage 2. The additional effect of shower energy reports, when combined with real-time feedback is identified by the comparison between the DUAL and RTF groups in IN stage 2. Differences between the effects of shower energy reports with and without real-time feedback identify policy interaction effects, i.e., whether the two interventions are substitutes or complements. Note that behavior in the CON group may not reflect a pure counterfactual, as subjects still receive a smart meter with temperature information as well as placebo emails, to hold experimenter interaction and Hawthorne effects fixed. We would underestimate the effects of our interventions to the degree that subjects respond to this by itself, but any relative comparison across intervention regimes would remain valid.

3.6. Behavioral predictions

In order to derive behavioral predictions for each of our experimental groups, we first briefly discuss the channels through which each of the two interventions is likely to work. Our theoretical framework shows that the effect of each regime depends on the degree to which it succeeds in overcoming the aggregate bias, which may be the product of multiple separate factors. Furthermore, real-time feedback and shower energy reports could be complements if they are relatively specialized and operate largely through different channels.

Real-time feedback visually displays live measurement of water use in the shower. This water volume information can debias individuals' beliefs about the amount of water they

use, but there is no additional information on energy use or CO_2 emissions due to water heating, so severe knowledge gaps about the environmental relevance of showering may remain. In addition, the steadily upward moving liter count is likely to significantly reduce inattention and self-control problems, as users are constantly facing the smart meter display, and the previously abstract and elusive notion of resource use suddenly becomes salient and palpable, infused with a sense of immediacy. It may also facilitate experimentation with various conservation strategies by keeping track of progress in real-time. As the RTF condition in our experiment is essentially a replication of the intervention by Tiefenbeck et al. (2018), albeit more minimalistic and in a sample without monetary incentives, we also expect to find comparable conservation effects:

Prediction 1. *Providing real-time feedback through the smart shower meter display in treatment RTF leads to a reduction in water and energy consumption in the shower.*

Shower energy reports provided personalized information about subjects' water use in the shower as well as additional information about energy use and CO_2 emissions. We therefore expect that the reports can help close knowledge gaps in these areas and thereby induce conservation behavior, since past evidence suggests that individuals tend to grossly underestimate the energy intensity associated with water heating (Attari et al., 2010). The second report also included a social comparison with a randomly assigned and anonymous peer, which might further add motivation (Allcott, 2011).

Prediction 2. *Providing information through shower energy reports in treatment SER leads to a reduction in water and energy consumption in the shower.*

As the shower energy reports are not immediately salient while showering, the effect of knowledge gains could be stifled by remaining barriers like limited attention or self-control problems that can be better targeted by real-time feedback.[15] Vice versa, the effect of real-time feedback may be attenuated if subjects remain unaware of the energy and carbon intensity of warm water use. If the two interventions indeed work largely through these separate behavioral mechanisms, a combined intervention should leverage all mechanisms at the same time. As we argue in Section 2 , shower energy reports and real-time feedback could therefore become complements in the sense that one intervention makes the other more effective when implemented jointly. Thus, we derive the following prediction:

Prediction 3. *Shower energy reports in IN stage 2 lead to a larger (marginal) reduction in water and energy consumption in the shower for subjects who also receive real-time feedback (DUAL group) than for subjects who do not receive real-time feedback (SER group).*

[15] In principle, it is possible that participants also become more attentive about resource use even without visual aid through the smart meter, as would be predicted by rational inattention models when updates in beliefs about environmental impacts are sufficiently large. However, if there is such an effect, it may prove short-lived once reports fade out of memory and resolutions cool off (Allcott and Rogers, 2014; Schwartz and Loewenstein, 2017).

Note that in the communication with the participants, the study was primarily framed as an *energy* conservation study, as water scarcity was less of an issue in Germany at that time. By contrast, energy conservation and the transformation of the energy sector ranked high on Germany's policy agenda. Nevertheless, in the subsequent analyses, we report results for both energy and water consumption. While greenhouse gas emissions arise mostly due to water heating, conserving water may be an objective in itself, especially given that climate change increases the likelihood of droughts and water stress even in parts of the world (like Germany) that previously have not suffered from water scarcity (European Environment Agency, 2021). As energy use is calculated as a product of water volume, temperature gradient, and constant factors (see 4.1 for details), the amount of energy and water consumed in the shower are highly correlated. Any change in water consumption (i.e., by shortening shower time or reducing the flow rate) affects energy consumption proportionally. The only margin of adjustment that has a different effect on these two outcomes are changes to the water temperature.

4. Data and descriptive statistics

4.1. Measurement data on resource use behavior

For every water extraction in the shower, the smart meters measured, among others, the volume of water used, its average temperature, and the average flow rate (i.e., volume per time unit). The amount of energy used was then calculated based on volume and temperature data, using the standard engineering formula for heat energy.[16] Every subject had a shower meter installed for the whole duration of the study, starting from early December 2016. At the end of the study, in early March 2017, we retrieved the devices and read out the data manually.[17] In this way, we were able to extract an initial data set of 21,469 showers by 327 participants. Unfortunately, no data could be obtained in 24 cases, either because the device was defective or because subjects never used it, or because subjects simply disappeared without a trace (and their shower meters with them).

A number of data cleaning steps are performed before running the empirical analyses. We briefly describe the most important steps here; a more detailed documentation can be found in Appendix C. First, we drop the very first data point of each participant, as they usually started with a test run to check if the device was working. Following Tiefenbeck et al. (2018), we further drop any water extraction with volume below 4.5 liters (in total 2,942 extractions), as these are unlikely to be actual showers but rather minor extractions for other purposes such as cleaning. As there are rare cases in which the device can pro-

[16] The formula for energy use of water heating is $Q = m \times c_p \times \Delta T$, with heat energy Q, mass of water m, heat capacity c_p , and ΔT the difference between the measured water temperature and cold water temperature (assumed to be 12 degrees Celsius). Following Tiefenbeck et al. (2018), we also assume boiler efficiency losses of 35% and distribution losses of 24%.

[17] We already started retrieving some devices in late February, but as the retrieval process was drawn out over a period several days, the end of the study was in early March for most subjects.

duce errors when storing data, we further remove 37 extreme outlier points, defined as such by being more than 4.5 times the subject-specific interquartile range away from the closest quartile.[18] We further exclude 1 device with generally erratic data, 5 devices with fewer than 10 recorded extractions, as well as 3 devices with an abnormally large baseline consumption of 168 liters or more per shower, which is about 40 liters (1.5 standard deviations) away from the rest of the field. In 8 cases, the integrated temperature sensor became defective after some time, and we impute missing information with the average temperature of showers taken while the sensor was still intact. The final data set used for our empirical analyses includes 17,942 showers by 318 participants.

The shower meter stores the temporal order of showers, so we can easily classify each shower into baseline or intervention stage, as real-time feedback (in the RTF and DUAL groups) started from the eleventh shower. Assigning showers to intervention stage 1 (pre-reports) or stage 2 (post-reports) is slightly trickier, as the device has no counter for global time. Fortunately, the smartphone app stores the date and time of each data upload, which allows us to construct time bounds for when a shower took place. Specifically, we know that a shower cannot have occurred after the time at which it was uploaded, and also not before the time of the last previous data batch, because otherwise it would have been uploaded then already. Combined with knowledge about the order of observations, we can assign approximate dates to each shower, assuming that the time that passes between one shower and the next remains roughly constant. For example, if three shower observations were uploaded at day t and the last previous upload occurred at day $t-3$, then we would assign the first of these showers to day $t-2$, the second to $t-1$, and the third shower to day t. We instructed subjects to use the app regularly starting from 11 Jan 2017 — two weeks before the first energy report (sent out at 2:30pm on Jan, 24th) —, and sent additional reminders before each energy report email (or placebo email) was sent out.

Using this timing information from data uploads, we classify observations into pre-report showers (IN stage 1) or post-report showers (IN stage 2). In particular, we know from mini-survey response data when subjects likely read the email and use this as cutoff date. For non-responders, we use the time at which we sent the email as cutoff date instead; the response rate for the first email was 82.7%.[19] Observations with upper time bound before the cutoff date are assigned to IN stage 1 and observations with lower time bound after the cutoff date are assigned to IN stage 2. For observations that fall into a range of uncertainty around the date on which the subject read/received the email, we intrapolate their dates based on the assumption that the frequency of showering was constant within that time range, and then use these intrapolated dates to assign them into

[18] We are particularly strict in only excluding the most implausible data points here. Conventionally, 1.5 or 3 times the interquartile range (IQR) are used as criterion for outliers. For a normal distribution, 4.5 times the IQR away from the nearest quartile corresponds to 6.745 standard deviations away from the mean.

[19] 47% of subjects responded to the mini-survey within the same day that the email was sent out, and 77% responded within one week.

the first or second intervention stage.[20]

A final complication comes from subjects who did not manage to upload any data to the app. For these non-uploaders, we impute the timing of shower energy reports based on the assumption that it follows the same distribution for uploaders and non-uploaders. To operationalize this, we use timing information from uploaders to estimate the probability that a shower took place after receiving the first (second) report for each shower, based on its temporal order, and then assign the implied post-report probabilities to showers of non-uploaders. Appendix Figure A3 plots the estimated CDFs, and more details on the imputation procedure are provided in Appendix D. We also consider alternative definitions of intervention periods for robustness checks.

4.2. Survey data

To supplement our behavioral data on resource use in the shower, we administered several questionnaires. In the baseline survey, we collected information on individual characteristics (i.e., age, gender, etc.), perceived water use in the shower, shower comfort (i.e., how much they enjoy showering), environmental attitudes and beliefs, as well as a number of personality attributes (i.e., Big Five, patience, etc). In the post-intervention survey, we again collected self-reported data on perceived water use, shower comfort, and environmental attitudes. Furthermore, we administered mini-surveys with each energy report, in which subjects were asked to estimate their resource use in the shower.

We mainly make use of information on water use perceptions, shower comfort, and environmental attitudes, and how they change in response to our interventions. Environmental attitude is elicited using four items about pro-environmental behavior and identity, e.g. "I do what is right for the environment, even when it costs more money or takes more time".[21] Shower comfort is elicited using five items on how much subjects enjoy showering, e.g. "I find it relaxing to take a shower".[22] We create indices for shower comfort and environmental attitude, respectively, by taking the simple average of the individual's responses to the relevant items (rated on a 4- or 5-point Likert scale) and then normalizing to mean 0 and standard deviation 1. For perceived water consumption, we asked subjects to estimate how many liters of water they typically use when taking a shower. These estimates can then be directly compared to their actual water use as measured by the smart meter. Note that we refrained from eliciting subjects' beliefs about

[20] For example, say we know for certain that shower s occurred at 8am on Jan, 22nd (pre-report), and shower $s+3$ occurred at 9am on Jan, 25th (post-report). This leaves the stage of showers $s+1$ and $s+2$ ambiguous. To assign these, we would assume that shower $s+1$ occurred in the morning of the 23rd and shower $s+2$ in the morning of the 24th, thus putting both showers before the first report, which was sent out at 4:30pm on the 24th.

[21] The other items are "Environmental friendliness is part of my personal identity", "How often do you try to conserve water?", and "How often do you try to conserve energy?". We also include a set of questions adapted from Nolan et al. (2008) in the baseline questionnaire.

[22] The other items are "I like showering", "For me, taking a shower is just a means to an end", "I like to let my mind wander when I shower", and "I try to shower as quickly as possible".

Table 1: Descriptive statistics – baseline showers

	Mean	Std. dev.	10th pctile	Median	90th pctile	Obs.
Energy use [kWh]	2.21	1.91	0.43	1.71	4.58	2503
Volume [liter]	37.77	30.40	9.30	29.60	75.70	2503
Duration [min]	6.99	5.00	1.97	5.82	13.00	2503
Temperature [Celsius]	36.14	5.23	32.00	37.00	40.00	2477
Flow rate [l/min]	5.71	2.45	2.80	5.40	9.10	2503

Includes only showers taken in the baseline stage, i.e., first 10 showers and before subjects read/received shower energy reports. For temperature statistics, devices with broken temperature sensors are excluded. Duration is net of any breaks and calculated by dividing water volume by flow rate.

energy use and carbon emissions from water heating, because we did not want to raise awareness about these issues and risk undermining the shower energy report treatments.

4.3. Sample characteristics and baseline behavior

All study participants were students at universities in Bonn or Cologne living in single-person dorm apartments, so our sample is rather homogeneous. From the 318 participants represented in our main dataset, 203 lived in a dorm in Bonn and 115 lived in a dorm in Cologne. The female share was 61 percent. Average age was 23.8 years (median 23 years), with students from all stages of their studies being represented in our sample.

Using the nine showers (the first being excluded) in the baseline stage, where only the current water temperature was displayed, we can construct measures of each subject's baseline resource use behavior. Table 1 presents descriptive statistics about baseline energy and water use per shower, as well as shower duration (net of breaks), water temperature, and flow rate. On average, showers in the baseline stage feature 7 minutes of water flow, which amounts to 37.77 liters of water. On average, water is heated up to a temperature of 36.14 degrees Celsius, resulting in energy consumption of 2.21 kWh per shower. There is substantial variation across showers, as observed from the standard deviations and different quantiles of the distributions. Water and energy consumption follow a right-skewed distribution, thus the median energy use per shower (1.71 kWh) is substantially lower than the mean. As the share of cold showers is extremely low in our sample (only 3.7% of showers have an average temperature of $21^{\circ}C$ or lower), water and energy usage is almost perfectly collinear, with a Pearson correlation coefficient of 0.9755. The average flow rate of 5.71 liters per minute is relatively low, likely due to dorm infrastructure not being up to modern standards — flow rates of 10-12 liters per minute are more typical for German households.

Table 2: Randomization checks and extensive margin responses

	Panel A. Baseline averages by individual					Panel B.
	Energy use [kWh]	Volume [liter]	Duration [min]	Temperature [Celsius]	Flow rate [l/min]	Number of showers
RTF group	-0.111	-1.253	0.284	0.086	-0.124	-2.312
	(0.215)	(3.427)	(0.597)	(0.595)	(0.370)	(5.183)
SER group	-0.077	-2.096	0.166	0.962	-0.441	3.393
	(0.218)	(3.431)	(0.547)	(0.608)	(0.319)	(5.226)
DUAL group	-0.071	-1.215	0.126	0.323	-0.151	3.224
	(0.226)	(3.571)	(0.578)	(0.556)	(0.357)	(5.861)
Constant	2.237	38.316	6.797	35.681	5.832	55.312
	(0.163)	(2.539)	(0.411)	(0.447)	(0.240)	(3.698)
Observations	316	316	316	314	316	318
R-squared	0.001	0.001	0.001	0.011	0.005	0.005
F-test: p-value	0.965	0.945	0.972	0.351	0.550	0.669

Robust standard errors in parentheses. The omitted category is the CON group. For two participants, the device was not able to record information on baseline showers, but we could extract valid data on showers in later stages; hence the number of observations is only 316 in most columns. In addition, two participants with initially defective temperature sensors are excluded in column 4.

4.4. Randomization checks

Our identification strategy relies on randomization producing treatment groups that are comparable with regard to observable and unobservable subject characteristics. Although it is naturally impossible to test the latter, we can check balance on observable baseline characteristics. Panel A of Table 2 shows results from regressing various measures of subjects' baseline behavior on assigned treatment groups. The differences between groups are very small and treatment assignment is insignificant for predicting any of the behavioral measures, so randomization seems to have worked well. We also check for balance along background characteristics and survey responses (see Table A1 in Appendix A), and again find that treatment assignment is statistically insignificant. Importantly, self-reported environmental attitude and shower comfort are comparable across groups.

4.5. Number of showers

On average, we observe 56.8 showers per individual over roughly 12 weeks of our study, which corresponds to a frequency of about two showers every three days. However, the net frequency (i.e., adjusting for absences) might be closer to one shower per day, as our study period included a two weeks Christmas break. In Panel B of Table 2, we check whether the number of showers per individual differs across experimental conditions, but we find that treatments have no effect on the number of showers ($p = 0.669$). Hence, our interventions do not seem to induce adjustments along the extensive margin, and we do not need to worry about subjects compensating shorter showers with more showers,

substituting behavior to other facilities (e.g. wash basin, gym showers), or about them compromising on basic hygiene needs. This means that we can make use of the full panel structure of our data and analyze (intensive-margin) water and energy conservation effects at the level of individual shower observations.

4.6. Presence of imperfect information and behavioral biases

Before moving on to the analysis of our experimental interventions, we provide suggestive evidence that individuals' resource consumption in our setting may indeed be subject to biases due to imperfect information and limited attention.

First, we make use of the pre-intervention questionnaire and compare subject's perceptions of their own water use per shower to their actual baseline water use as measured by the smart meter. Figure 5 shows that subjects' estimates are all over the place; we cannot even reject the null hypothesis that estimated and measured water use are uncorrelated (Pearson's $\rho = 0.0925, p = 0.1308$). This demonstrates that subjects lack information about their own behavioral outcomes prior to any intervention.[23] Interestingly, the mean estimate (43.4 liters) and median estimate (30 liter) across subjects were not too far from the typical baseline water usage per shower in our sample. This is reminiscent of a "wisdom of crowds" phenomenon and suggests that, on average, our interventions should not work through debiasing beliefs about water use.

However, people may be particularly unaware about the link between water heating for showering and energy consumption, and hence CO_2 emissions. For example, Attari et al. (2010) show that consumers are in general highly prone to underestimating the amount of energy required for heating up water (e.g., water boilers, dishwashers). We did not elicit beliefs about energy intensity or carbon emissions in the original experimental sample, to avoid the risk of undermining our shower energy report treatments. We did, however, elicit beliefs about carbon emissions in a different sample of students living in the same dormitories three years after the original study ($n = 329$). For more details on this supplementary study, see Appendix E. Without additional information, these students underestimated the carbon impact of warm water use in the shower by a factor of 8 to 9 on average, even though the average guess for the amount of water used per shower was fairly unbiased. On average, students estimated that a typical shower causes emissions of 91.3 grams of CO_2 (median 35 grams), whereas the actual emissions amount based on the data from our main experiment is about 800 grams.[24] Thus, there might be a large potential for encouraging energy conservation through the information provided in shower energy reports (Byrne et al., 2018).

[23] We excluded 35 subjects who responded to the baseline survey more than 2 weeks after we distributed shower meters, as they have likely reached the intervention stage by then. We also exclude 3 outliers with estimates above 200 liters. The corresponding regression results are presented in Appendix Table A13.

[24] The average guess for amount of water used per shower was 40.4 liters. The survey was conducted in Nov/Dec 2019 among 329 residents of the exact same student dorms in which the original study took place in 2016/17. Only 4 surveyees had already participated in the original study.

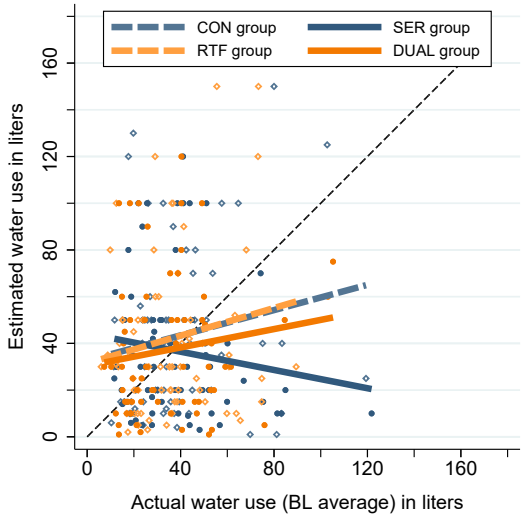


Figure 5: Pre-intervention awareness about water use per shower

Notes. This figure compares estimated water use from the baseline survey with actual water use in the baseline stage (showers 2 to 10), excluding late survey responders. 3 outliers with estimates between 200 and 600 liters are excluded. Point clouds consist of individual observations (hollow diamonds for CON and RTF, solid circles for SER and DUAL) and lines represent separate regression fits for each treatment group. The dashed line starting from the origin is the 45 degree line.

Although anecdotally compelling, finding direct evidence for inattention or self-control problems in the shower is trickier. The closest proxy we have is a baseline survey item on how much subjects agree with the statement "I like to let my mind wander when I shower.". 59% of our sample agreed or strongly agreed to the statement (25% strongly agree), whereas only 18% of subjects disagreed (13% weakly disagree, 5% strongly disagree). Moreover, subjects' response to this item is significantly correlated with their average baseline energy use in the shower (Pearson's $\rho = 0.1645$, $p = 0.004$). In fact, it is the single most predictive item for baseline consumption in the entire survey based on simple linear regressions. Our interventions could thus help reduce energy use by reminding subjects to stay focused and not lose track of time completely under the shower.

5. Estimation approach

Next, we describe our strategy for estimating the effects of our interventions on resource use in the shower. The empirical results will be presented in the following section.

5.1. Basic identification and estimation strategy

To formally estimate the effects of different intervention regimes, we exploit the randomized assignment of subjects into experimental conditions as well as the staggered introduction of real-time feedback and shower energy reports, which gives us a double-layered difference-in-differences setup. The differential changes in consumption behav-

ior across conditions from baseline stage to intervention stage 1 identify the causal effect of real-time feedback (RTF/DUAL versus CON/SER), and the additional changes from intervention stage 1 to stage 2 identify the causal effect of shower energy reports, both in isolation (SER versus CON) and in conjunction with real-time feedback (DUAL versus RTF). In this setup, interaction effects between the two interventions can be identified by comparing the incremental effects of shower energy reports with real-time feedback (DUAL) and without real-time feedback (SER).

If one was only interested in estimating the effect of real-time feedback in isolation, the most straightforward approach would be to simply compare how subjects in the RTF and CON groups change their behavior from the baseline stage to the entire intervention stage, without any need to consider shower energy reports or stage 2. To jointly estimate the effects of each intervention regime, using data from all experimental condition, we instead consider the following regression equation:

$$\begin{aligned} y_{it} \;=\; \alpha_i \;+\; IN_{it} \times \Big(\beta_0 \;+\; \beta_1 T_i^{R/D} \;+\; \beta_2 T_i^{S} \;+\; \beta_3 T_i^{D}\Big) \\ +\; IN_{it}^{s2} \times \Big(\gamma_0 \;+\; \gamma_1 T_i^{R/D} \;+\; \gamma_2 T_i^{S} \;+\; \gamma_3 T_i^{D}\Big) \;+\; \varepsilon_{it}\,, \end{aligned} \tag{8}$$

where the outcome variable y_{it} is energy (water) usage by individual i for shower number t, and α_i is the individual fixed effect. $T_i^{R/D}$, T_i^{D} and T_i^{S} are treatment group indicators, where superscript R/D denotes the combined real-time feedback groups RTF and DUAL, D denotes the DUAL group only, and S denotes the SER group only. Finally, IN_{it} is an indicator that takes the value 1 if observation it falls into the intervention stage ($t > 10$), and IN_{it}^{s2} is an indicator for showers that fall into intervention stage 2. As IN_{it} applies to the entire intervention period, IN_{it}^{s2} captures incremental changes in consumption from intervention stage 1 (pre-report) to stage 2 (post-report). Note that the stage 2 coefficients combine the effects of two distinct reports, the second containing also social comparison.

Given our formulation of the statistical model, we can interpret β_1 as treatment effect of real-time feedback on energy (water) use per shower in the first stage of the study, and γ_1 is its change in the second stage. This allows us to test *Prediction 1*. The relevant comparison for *Prediction 2* on the effect of shower energy reports in isolation is between SER and CON after the reports, which is captured by γ_2. Finally, γ_3 captures the marginal effect of adding shower energy reports to real-time feedback, by comparing DUAL and RTF in intervention stage 2. Finally, the comparison between γ_2 and γ_3 nails down the interaction effect between the two interventions and thus allows us to test for any potential complementarities (*Prediction 3*). As each test is associated with a separate hypothesis, we do not adjust our inference for multiple hypothesis testing (Rubin, 2021).

5.2. Estimating treatment effects on the treated

One complication in estimating the effect of shower energy reports is that 28% of subjects did not succeed in uploading any data to the Amphiro smartphone app before we sent out the reports, mostly due to technical problems (e.g., Bluetooth connection failure).[25] For these "non-uploaders", we were unable to provide informative shower energy reports. As the emails were generated automatically, non-uploaders in SER and DUAL groups received report templates with blanks where it was supposed to show statistics on resource use and environmental impacts. Effectively, this leads to imperfect treatment take-up of shower energy reports, although being less the result of deliberate non-compliance than unfortunate circumstances. For participants in the CON and RTF groups, it is inconsequential whether they successfully uploaded data.

To test Predictions 2 and 3 from Section 3.6, one possible approach to estimate treatment effects under imperfect treatment take-up would be simply to run an intention-to-treat (ITT) analysis, which only uses treatment assignment information and ignores that some subjects did not actually receive informative shower energy reports. While this would be the relevant parameter in many policy evaluation contexts, one of our main aims is to test whether the effect of receiving information on the energy use and carbon emissions due to hot water consumption interacts with the effect of receiving real-time feedback in the shower (Prediction 3). This would shed light on the potential importance of multiple biases for complementarites between behavioral interventions that we identify theoretically in Section 2. However, a stringent empirical test of this requires that subjects indeed have the opportunity to gain knowledge through shower energy reports. Thus, the more policy-relevant parameters in our case are the treatment effects on the treated (TOT) in the DUAL and SER groups, i.e., the effects of shower energy reports on subjects who managed to upload data prior to the report and thus received actual information on the environmental impact of their hot water consumption in the shower.

The first way in which we estimate the TOT is by simply comparing only the uploaders in SER and DUAL groups with subjects in the CON and RTF groups. The usual concern at this point would be that treatment take-up is not random. Fortunately, our setting limits potential endogeneity concerns for three reasons. First, we include individual fixed effects, so our estimates would still be unbiased if differences between uploaders and non-uploaders do not interact with the treatment effect. Second, subjects only knew that they should use the smartphone app to upload data, but we did not announce that we would use this data to construct shower energy reports. Thirdly, the main cause for non-compliance is not the lack of willingness to use the smartphone app, but unexpected technical failure, which is unlikely to be selected on the trend. To alleviate the most blatant endogeneity issue, we also exclude non-uploaders in the CON and RTF groups who did not report any technical problems. Appendix Tables A2 and A3 present additional

[25] Out of the 90 non-uploaders in our estimation sample, 63 have explicitly contacted us for technical problems encountered during their upload attempts.

balance checks for the TOT subsample and show that the experimental groups remain balanced along baseline characteristics.

The second way in which we estimate the TOT is by using random treatment assignment as instrument for actual take-up.[26] This can be shown to identify the so-called local average treatment effect (LATE), i.e., the average treatment effect for the sub-population of compliers, in our case the uploaders (Imbens and Angrist, 1994).[27] Compared to the "uploaders-only"-approach, the instrumental variables approach is consistent under weaker assumptions, but potentially inefficient. We will report the results from both TOT-approaches, but the estimates are very similar, suggesting that non-compliance due to technical issues was likely uncorrelated with conservation intentions in our sample.

6. Main empirical results

6.1. Average treatment effects

We start by presenting descriptive evidence on the resource conservation effects of our interventions in Figure 6 by plotting subjects' average changes in energy and water consumption per shower in intervention stage 1 (pre-report) and intervention stage 2 (post-report) compared to the baseline period. The differences-in-differences across treatment groups then correspond to the average treatment effects. Note that the stage 2 averages need to interpreted as combining effects of two distinct reports, one without and one with social comparison. In order to show the treatment effects on the treated (TOT), i.e., the effect of *informative* shower energy reports, we use the uploaders-only approach of excluding non-compliers in SER and DUAL as well as non-compliers without technical problems in CON and RTF.

Figure 6 essentially summarizes our main results in eight bars. The patterns are very similar for energy and water consumption. The four bars to the left of the dashed vertical line represent the change in resource use per shower in intervention stage 1 compared to the baseline stage. We can see that relative to subjects in the CON and SER groups, subjects in the RTF and DUAL groups with real-time feedback reduced their consumption drastically, by almost 0.4 kWh of energy and 6 liters of water per shower. Recall that there were no shower energy reports yet at this point. The four bars to the right of the dashed vertical line represent the change in energy use per shower from baseline stage to intervention stage 2, after shower energy reports were sent out. The first observation is

[26] To do this, we create new treatment indicators for the DUAL and SER groups that took the value 1 for showers in IN stage 2 by subjects who were assigned to the respective group *and* who uploaded data through the smartphone app that we could use to construct their shower energy reports. The previously defined ITT indicators are then used as instruments for these new indicators for receiving actual shower energy reports.

[27] This identification result holds under the condition that there are no "defiers", subjects who always do the opposite of what they are prescribed. This monotonicity condition holds by design in our study, because we control the eligibility of shower energy report treatment, so any participant in the sample can be classified either as complier or as never taker in the LATE framework.

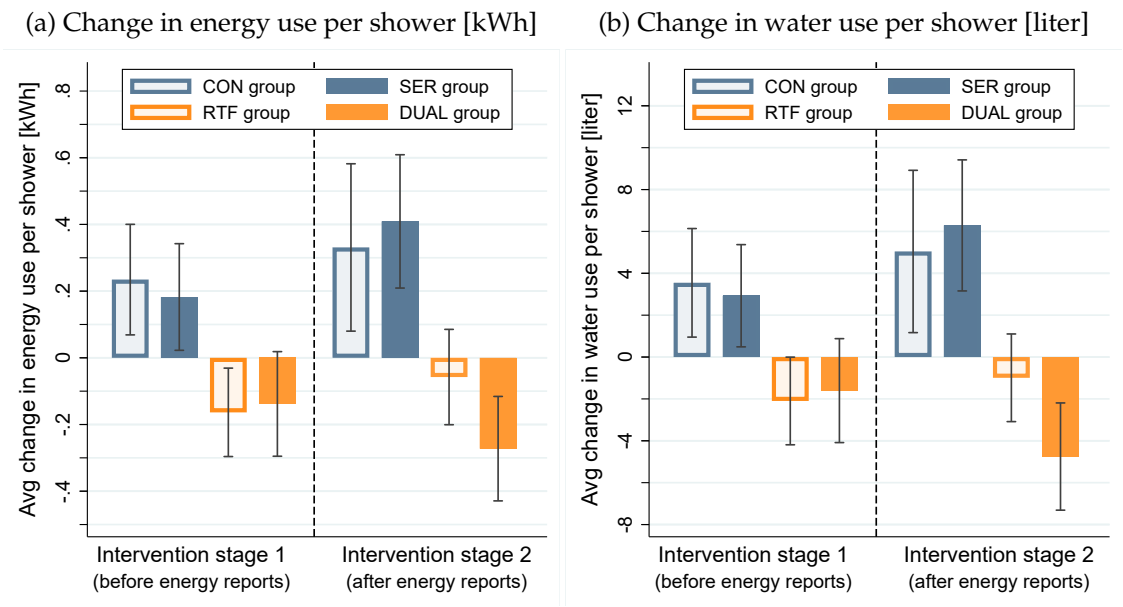


Figure 6: Descriptive evidence on resource conservation effects

Notes. The bars represent changes in average energy use and water use per shower compared to the baseline period. The error whiskers represent 90% confidence intervals. Non-uploaders in SER and DUAL as well as non-uploaders without technical problems in CON and RTF are excluded.

that average resource use in the control group further increased, which could be driven by weather effects, by pending exams leaving students stressed and in need for a long and warm shower, or by Hawthorne effects that decrease over time (Tiefenbeck, 2016).[28] The second observation is that the RTF group and the CON group followed a more or less parallel trend from intervention stage 1 to stage 2, hence the effect of real-time feedback in isolation remains nearly constant. The third observation is that providing shower energy reports in isolation does not seem to result in effective behavioral change: consumption of subjects in the SER group followed the CON group in close synchronization. In light of this, the fourth and final observation is particularly striking: shower energy reports are highly effective when combined with real-time feedback. In fact, subjects in the DUAL group are the only ones to defy the general upward trend and reduce their consumption considerably compared to subjects in the RTF group.

The descriptive evidence presented in Figure 6 is confirmed by formal empirical estimates based on the empirical strategy outlined in the previous section. Table 3 presents the regression results from estimating equation 8, with columns 1-2 using the uploaders-only approach, and columns 3-4 using the alternative LATE approach. To ensure that our statistical inference procedure is robust to arbitrary temporal interdependence of showers taken by the same person, we cluster all standard errors at the individual level. Appendix Figures A4-A5 show that all statistical test results are virtually identical when

[28] While the baseline phase fell mainly into an unusually warm and dry December, the main intervention months of January and February saw much higher precipitation. Exam periods at the universities began in mid-February.

using randomization-based inference methods (Young, 2019).

Focusing first on the effects of real-time feedback in isolation, LATE is preferable as it utilizes the full sample. We document a conservation effect of around 0.37kWh energy and 5.5 liters of water per shower from intervention stage 1 onwards (coefficient β_1). The effect does not change significantly in intervention stage 2 (coefficient γ_1); if anything, it becomes slightly stronger. Another direct way to estimate the effect of real-time feedback that is easier to interpret is to only compare subjects in the RTF and CON groups, since there is no need to take into account effects of shower energy reports. Appendix Table A4 shows that real-time feedback in isolation reduces resource use by 0.4 kWh of energy and 6.3 liters of water per shower compared to the CON group over the entire intervention period, which corresponds to about 17-18% of average baseline consumption.

Result 1. *Real-time feedback (in isolation) through the smart meter display led to a reduction in energy (water) consumption by about 0.4kWh (6.3 liters) or 17-18% per shower.*

For the shower energy reports, we need to focus instead on intervention stage 2 and account for imperfect compliance, to estimate the effects of actually receiving information on resource use and environmental impacts (see Appendix Table A4 for the intention-to-treat estimates). Recall that while the LATE approach is consistent even under strong endogeneity of treatment take-up, the uploaders-only approach is potentially more efficient and still consistent if technical issues in uploading data are as good as random. Table 3 shows that the point estimates obtained both approaches are very similar, implying that endogeneity of treatment take-up is likely not a major issue in our sample, whereas the standard errors are smaller in the uploaders-only approach. Contrary to prediction 1, shower energy reports in isolation had no significant conservation effect in the SER group (coefficient γ_2), and the point estimates even go in the opposite direction. While the null effect is not very tightly estimated, we can rule out reductions of greater than 4-5% (0.08kWh and 1.95 liters per shower) with 90% confidence in our preferred uploaders-only specification. This would be consistent with effect sizes in the order of magnitude found in previous studies (e.g., Allcott, 2011). Note that lower statistical power due to non-compliance does not explain the insignificant coefficient for SER compared to the significant coefficient for RTF, since the standard errors for β_1 and γ_2 are similar. We can also reject the hypothesis that shower energy reports in isolation were as effective as real-time feedback in isolation ($p = 0.009$).

Result 2. *Shower energy reports in isolation did not induce any significant reduction in energy and water consumption per shower.*

In stark contrast, we find that adding shower energy reports in the DUAL group induced subjects to further reduce their consumption by around 0.23 kWh of energy and 3.8 liters of water per shower in intervention stage 2 (coefficient γ_3), corresponding to another 10%p reduction from baseline consumption and about 60% of the effect of real-

Table 3: Treatment on the treated (TOT) estimates

	Uploaders-only		*LATE*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
(β_0) Intervention	0.207*	3.067*	0.200**	2.961*
	(0.108)	(1.663)	(0.099)	(1.530)
(β_1) Intervention × RTF/DUAL	-0.388***	-5.762***	-0.368***	-5.514***
	(0.130)	(2.071)	(0.122)	(1.932)
(β_2) Intervention × SER	0.013	0.543	0.003	0.476
	(0.151)	(2.361)	(0.132)	(2.051)
(β_3) Intervention × DUAL	0.031	0.524	0.109	2.140
	(0.111)	(1.808)	(0.106)	(1.719)
(γ_0) IN stage 2	0.110	2.191*	0.152*	2.756**
	(0.085)	(1.319)	(0.090)	(1.360)
(γ_1) IN stage 2 × RTF/DUAL	-0.023	-1.234	-0.054	-1.550
	(0.109)	(1.805)	(0.113)	(1.806)
(γ_2) IN stage 2 × SER	0.124	1.336	0.079	0.573
	(0.124)	(1.993)	(0.149)	(2.326)
(γ_3) IN stage 2 × DUAL	-0.230**	-3.836**	-0.226*	-4.013*
	(0.109)	(1.908)	(0.122)	(2.152)
(α_i) Individual fixed effects	*yes*	*yes*	*yes*	*yes*
p-value: $\gamma_2 = \beta_1$	0.007	0.018	0.026	0.053
p-value: $\gamma_2 = \gamma_3$	0.033	0.062	0.114	0.149
Clusters	261	261	318	318
Observations	14712	14712	17942	17942
R^2	0.412	0.415	0.004	0.004

In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported R^2 in columns (3) and (4) is the within R^2. Standard errors in parentheses are clustered at the individual level. Permutation-based inference procedures are presented in Figure A4 and A5. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

time feedback in isolation. Thus, information on environmental impacts of hot water consumption was not ineffective per se in our setting, but in fact boosted conservation efforts considerably when administered in combination with real-time feedback on water usage via the smart meter displays. This contrast between individuals' responses to shower energy reports with and without real-time feedback is all the more remarkable given that in the former case they had already cut their consumption significantly in intervention stage 1, thus leaving less room for further behavioral adjustments. The exact hypothesis for complementarity further requires to test not only whether the incremental consumption reduction in the DUAL group was different from zero, but also from the effect in the SER group (i.e., whether $\gamma_3 = \gamma_2$). This difference is statistically significant in the uploaders-only specification ($p = 0.033$), although in the less efficient LATE-specification it would be weakly significant only when using a one-sided test.

Result 3. *Combining real-time feedback with shower energy reports further reduced energy (water) use by around* 0.23 *kWh (*3.8 *liters) per shower, in addition to the conservation effect of real-time feedback in isolation.*

Overall, we observe a sizable complementarity between the two interventions in our setting. This is consistent with our theoretical framework, which shows that in the presence of multiple biases, behavioral interventions may need to address all significant sources of bias simultaneously in order to unfold their full effect. While shower energy reports provide information about resource use and associated environmental impacts, conservation efforts may be hindered by residual biases such as lack of salience. Real-time feedback through smart meters could thus turn environmental considerations into action by keeping them on top of people's mind in the heat of the moment. We will analyze the underlying mechanisms more closely in Section 7.

6.2. Robustness checks on timing

As the timing of observations with regard to intervention stage 2 involves a degree of fuzziness for subjects who did not use the app frequently, we conduct a number of robustness checks. First, we use a donut hole approach that excludes the around 10% of shower observations with the highest uncertainty about whether they occurred before or after the first shower energy report; Appendix Table A5 shows that our results remain nearly unchanged.[29] Second, our results are also robust to using an alternative definition of report timing for non-uploaders that deterministically assigns non-uploaders into intervention stage 2 based on the median study completion value among uploaders rather than the full cumulative distribution (see Appendix Table A6). Finally, we estimate a specification in which we assign all subjects into intervention stage 2 based on when we sent out the first shower energy report email, rather than based on when they actually read the reports (proxied by mini-survey response date). Appendix Table A7 shows that our results are robust to using this alternative timing indicator.

6.3. Margins of behavioral adjustment

Subjects could conserve energy by reducing the temperature to which water is heated to and the overall amount of water that needs to be heated up. Appendix Table A8 shows that reductions in water temperature seem to be at most a minor factor in our sample,

[29] To be more precise, we calculate for each shower a probability that it occurred after reading the first shower energy report (or placebo email). Notice that for many showers, these probabilities are either 0 or 1, because they were uploaded before the report or after the first post-report upload, respectively. For observations within the range of uncertainty, we calculate approximate probabilities assuming that the frequency of showering is constant. We then exclude all observations with probability between 10% and 90%, i.e., those with significant uncertainty about whether they occurred before or after the report. For non-uploaders, we use the same cutoffs of 10% and 90%, but based on the CDF for uploaders (see Figure A3).

perhaps for hedonic reasons.[30] Hence, water and energy usage tend to be very closely aligned with each other, and energy conservation effects are almost equivalent to water conservation effects, in relative terms. Water conservation, in turn, can be achieved by adjusting time spent under the shower, water flow rate (i.e., liters of water per minute), and the covariance structure. The data suggests that subjects respond to the interventions mostly by taking shorter showers and reducing the flow rate during long showers.

6.4. Treatment effect dynamics

In a next step, we investigate how resource conservation outcomes changed over time, with a focus on the last 5-6 weeks period of our study, after the first energy reports were sent out (IN stage 2). This allows us to test whether effects declined over time or remained stable and whether the second shower energy report (containing social comparison) may have induced behavioral responses. To estimate effect dynamics, we extend the empirical model from equation 8 by interactions with a time variable Z_i:

$$\begin{aligned} y_{it} = \alpha_i &+ IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D} + \beta_2 T_i^S + \beta_3 T_i^D\right) \\ &+ IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^S + \gamma_3 T_i^D\right) \\ &+ IN_{it}^{s2} \times Z_i \times \left(\delta_0 + \delta_1 T_i^{R/D} + \delta_2 T_i^S + \delta_3 T_i^D\right) + \varepsilon_{it}\,. \end{aligned} \tag{9}$$

We explore two variants of Z_i. In the first variant, we look additionally at energy use per shower after the second shower energy report, which was sent about two weeks after the first report. In the second variant, we interact each treatment group indicator with a linear time trend, so the δ coefficients can be interpreted as weekly depreciation (or appreciation) rate of energy conservation effects by intervention regime.

Table 4 suggests that the effect of shower energy reports in the DUAL group seemed to gradually unfold over time. The point estimates in columns (1) and (2) indicate that the average conservation effect is driven largely by the final 3-4 weeks of the study, after the second reports were sent out. However, this does not seem stem from a discrete jump, but rather from a continuous trend. In columns (3) and (4), we estimate that the conservation effect per shower in the DUAL group increases by a rate of around 0.08-0.09 kWh every week. These descriptive results need to be interpreted with caution, as the relevant coefficients are statistically insignificant. However, we note that the point estimates for shower energy reports in isolation (SER group) show no signs of any quantitatively significant dynamic pattern. The effect of real-time feedback in isolation also appears to stay constant in intervention stage 2, overall showing no signs of weakening within the 3

[30] At 40 liters and a base temperature of $37°C$, reducing energy conservation by 0.1kWh would require lowering the temperature by more than $1°C$, ceteris paribus.

Table 4: Treatment effect dynamics

	$Z_i = \mathbb{I}\{post\ report\ 2\}$		$Z_i =$ # *weeks after report 1*	
	(1) Uploaders	(2) LATE	(3) Uploaders	(4) LATE
...	...	...	...	...
(γ_0) IN stage 2	0.091	0.135	0.016	0.066
	(0.095)	(0.103)	(0.108)	(0.115)
(γ_1) IN stage 2 × RTF/DUAL	-0.038	-0.059	0.037	0.016
	(0.120)	(0.127)	(0.149)	(0.154)
(γ_2) IN stage 2 × SER	0.149	0.102	0.210	0.163
	(0.135)	(0.156)	(0.151)	(0.178)
(γ_3) IN stage 2 × DUAL	-0.086	-0.068	0.009	0.049
	(0.109)	(0.120)	(0.154)	(0.169)
(δ_0) IN stage 2 ×Z_i	0.032	0.029	0.032	0.029
	(0.090)	(0.087)	(0.023)	(0.022)
(δ_1) IN stage 2 × RTF/DUAL ×Z_i	0.025	0.009	-0.021	-0.024
	(0.128)	(0.125)	(0.035)	(0.034)
(δ_2) IN stage 2 × SER ×Z_i	-0.043	-0.040	-0.030	-0.028
	(0.123)	(0.133)	(0.036)	(0.041)
(δ_3) IN stage 2 × DUAL ×Z_i	-0.245	-0.273	-0.082	-0.095
	(0.201)	(0.207)	(0.055)	(0.058)
(β_j) Intervention indicators	*yes*	*yes*	*yes*	*yes*
(α_i) Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Clusters	261	318	261	318
Observations	14712	17942	14712	17942
R^2	0.413	0.004	0.413	0.005

The results are obtained by estimating equation (9). The full table with all the coefficients is presented in Table A9. In columns (1) and (3), we exclude all non-uploaders in SER and DUAL, as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (2) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported R^2 in Columns (2) and (4) is the within R^2. Standard errors in parentheses are clustered at the individual level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

months of our experiment.[31]

There are several potential behavioral explanations for this descriptive pattern of increasing responses over time that we observe in the DUAL group.[32] The first explanation is that the peer comparison element in the second report provided important additional social motivation to conserve energy, which then interacted with real-time feedback; this would be in line with our theoretical framework as well as previous literature. A sec-

[31] This is consistent with other interventions using smart shower meters (Agarwal et al., 2020; Byrne et al., 2022). Energy conservation studies in other settings show some degree of backsliding over time after being exposed to non-monetary interventions (e.g. Allcott and Rogers, 2014; Ito et al., 2018), although investments into physical capital may alleviate this issue in the long term (Brandon et al., 2017).

[32] Another potential explanation is that the apparent increase in estimated effects over time is a statistical mirage driven by decreasing measurement error about when subjects were treated with shower energy reports.

ond explanation is that subjects may have required some time to discover new strategies for further reducing resource use. This experimentation channel is consistent with Appendix Table A8, which suggests that subjects in the DUAL group conserved resources in the second intervention stage by reducing the flow rate overproportionately during longer showers. To further investigate this channel, Appendix Table A10 reports a specification that allows for discontinuities as well as differential trends after each report; although admittedly noisy, the estimates suggest a more or less constant and continuous (downward) slope of resource consumption in the DUAL condition that already begins starting from the first report. This suggests that the effects are unlikely to be driven by social comparison alone, although it may have played a role in reinforcing them. Importantly, the results speak against pure Hawthorne effects or short-lived attention boosts, as these would rather predict an "action-and-backsliding" pattern (Allcott and Rogers, 2014; Schwartz and Loewenstein, 2017).

6.5. Heterogeneity by baseline consumption

A frequent finding in the literature is that households or individuals with high baseline consumption tend to respond more strongly to interventions that foster conservation behavior (e.g., Allcott 2011; Ferraro and Price 2013; Tiefenbeck et al. 2018). Policy makers could therefore improve cost-effectiveness by targeting high-baseline consumers. To estimate heterogeneity by baseline energy use, we extend the statistical model in equation (8) by adding interactions with baseline consumption, measured using subject's average energy usage per shower in the baseline stage. Alternatively, we estimate a specification where we interact with an above-median indicator. Appendix Table A11 shows that, consistent with previous studies, the effect of real-time feedback increases with baseline usage. In intervention stage 2, subjects with 1 kWh higher baseline in the RTF group reduced their energy use by an additional 0.25 kWh ($p = 0.083$) on average, and above-median baseline users (mean 3.30 kWh) saved 0.63 kWh ($p = 0.043$) of energy more per shower compared to subjects with below-median baseline use (mean 1.17 kWh). It also appears that providing information through shower energy reports in the DUAL condition was about twice as effective for above-median users compared to below-median baseline users, although the difference is not statistically significant, whereas shower energy reports in isolation (SER) had no significant effects in either subpopulation.

7. Underlying mechanisms

The empirical results show that, in our setting, shower energy reports appeared to be ineffective in isolation, but induced large and significant conservation effects when combined with real-time feedback, which suggests that our interventions were complements. Through the lens of the theoretical framework in section 2, our proposed explanation for

this finding is that the two interventions targeted separated behavioral biases. Shower energy reports may have increased knowledge about environmental impacts of warm water use in the shower, but this in itself may not achieve reductions in energy consumption if subjects still suffer from limited attention or self-control problems. Real-time feedback could help mitigating these problems and thus enable knowledge gains to translate into conservation behavior. If, on the other hand, shower energy reports and real-time feedback both operated through the same mechanisms, we would generally not expect complementarities unless through some type of crowding in effect, e.g., if the combined intervention leads to positive attention or motivation spillovers. In this section, we conduct a number of analyses to explore the mechanisms underlying our results.

7.1. Awareness about resource intensity and environmental impacts

A crucial element of both interventions in our study is that they can enable learning about the outcomes of one's behavior. Real-time feedback through the smart meter provides immediate display of water use (and temperature) for the current shower. Shower energy reports also contain information of individuals' entire history of water (and energy) use per shower since the start of the study, with the difference that it comes in retrospect. Hence, a first manipulation check for our interventions is to analyze their effect on subjects' awareness about their own water use per shower.

In the post-intervention survey at the end of the study, we asked subjects to again estimate the amount of water they typically use per shower. Recall that prior to the interventions, subjects' assessments were virtually uncorrelated with their actual water use (see Figure 5). This picture changes after the interventions. Figure 7 plots individuals' post-intervention estimates as a function of their average water use per shower as measured by the smart meter. The corresponding regression table A13 is presented in Appendix A. Whereas subjects in the CON group remained as ignorant as before, the estimates by subjects who received real-time feedback (RTF and DUAL group) were much more aligned with their actual consumption patterns, as indicated by the fitted regression lines moving closer to the identity line. Importantly, shower energy reports in isolation (SER group) also induced significant learning effects about water use compared to the control group ($p = 0.039$). Moreover, we cannot reject that the learning slopes among uploaders are different between SER and DUAL ($p = 0.522$). We obtain similar results when we focus instead on the magnitude of estimation errors as outcome variable. Table A14 in Appendix A shows that subjects estimation errors in the three treated groups are on average about 20-30 percentage points closer to their actual water use than subjects in the CON group, and notably, the effect is virtually the same for SER, RTF, and DUAL groups.

Taken together, the results show that subjects in our study became better informed about their own consumption behavior in the shower through our feedback interventions. However, belief updating about water usage alone cannot explain our main re-

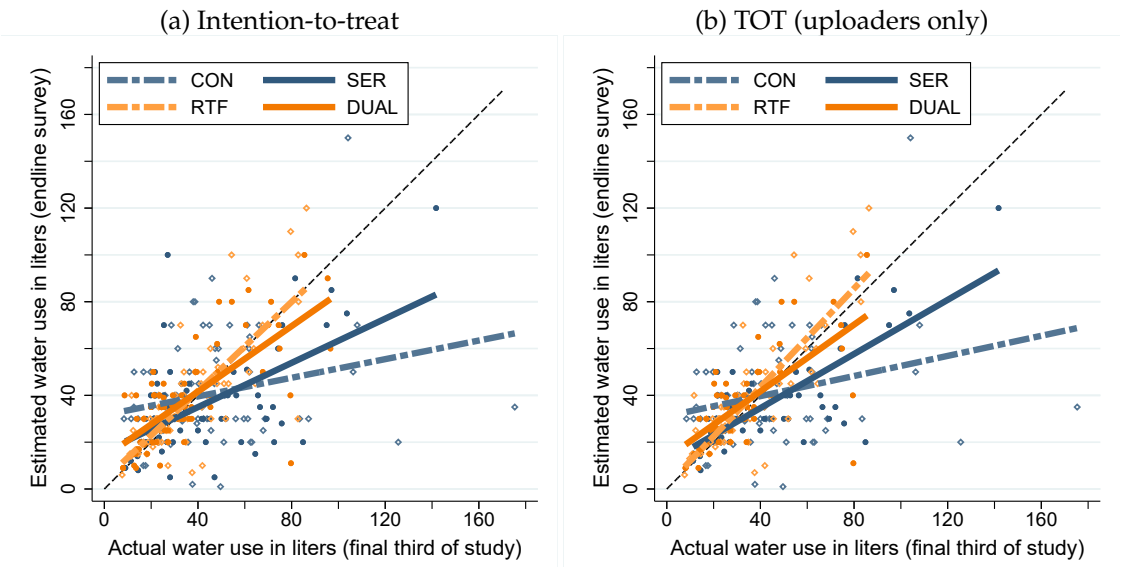


Figure 7: Post-intervention awareness about water use per shower

Notes. Both graphs compare subject's estimates in the final questionnaire with their measured average water use for the last third of shower observations. Graph (b) only uses the subsample defined for the uploaders-only approach. 7 outliers with estimates between 200 and 500 liters are excluded. Point clouds consist of individual observations (hollow diamonds for CON and RTF, solid circles for SER and DUAL) and lines represent separate regression fits for each treatment group. The dashed line starting at the origin is the 45 degree line.

sults. First, subjects' prior beliefs about water use were by and large unbiased even in the control group: on average, low-baseline users overestimated and high-baseline users underestimated. Second, we observe signficant belief updating in the SER group that does not translate into resource conservation effects. This points to the importance of the immediacy and salience of the real-time feedback intervention, which can help subjects track their water use while showering and overcome inattention problems.

In contrast to real-time feedback, shower energy reports did not only contain information about water, but also on energy usage due to water heating and environmental impacts in terms of CO_2 emissions. This can explain why subjects in the DUAL group reduced their energy consumption even further after receiving the reports. As a manipulation check for whether subjects responded to this information, we conducted a supplementary survey in a new sample of 329 students at the end of 2019 (see also Section 4.6). After eliciting prior beliefs about water consumption and CO_2 emissions per shower, we randomly presented one fact sheet (out of three) to each surveyee, mimicking the basic informational content of our original interventions. The "CON sheet" only reported the average water temperature in the shower, the "RTF sheet" also included the average amount of water used, and the "SER sheet" further added information on energy use and CO_2 emissions. After presenting the fact sheets, we elicited posterior beliefs as well as conservation intentions. We find that, relative to RTF sheet, surveyees that received the SER sheet information drastically adjusted their beliefs about CO_2 emissions upwards ($p < 0.001$), and their self-reported intention to take shorter showers in the fu-

ture increased by with a 0.24 standard deviations ($p = 0.003$). For further details on the supplementary survey, see Appendix E.

Thus, the shower energy reports increased knowledge about the environmental impact of hot water usage in the shower as well as conservation intentions, yet they were only associated with significant conservation effects when combined with real-time feedback. A key insights of our theoretical framework is that if biased behavior arises from multiple different sources, a narrowly-targeted intervention can be undermined by residual biases (Anna Karenina effect). Hence, a likely explanation of our results is that, when shower energy reports were implemented in isolation, additional biases due to, e.g., limited attention or self-control problems have prevented knowledge gains and good intentions from translating into actual behavior.

7.2. Engagement with shower energy reports

One potential alternative channel is differential treatment engagement, in the sense that subjects across experimental conditions may have paid more or less attention to the interventions per se. For example, if previous exposure to real-time feedback induced subjects in the DUAL group to read shower energy reports more carefully than subjects in the SER group, this might lead to complementarity between the two interventions through some type of crowding in or foot-in-the-door effect.[33]

While the previous analyses show that shower energy reports induced significant learning effects also in the absence of real-time feedback, we can also compare engagement with the reports between SER and the DUAL groups more directly by making use of the mini-surveys that were attached to the reports. As described before, each email included a link to a survey in which we asked subjects to estimate their water usage per shower. The survey link was at the bottom of the email, so subjects had to scroll through all the statistics on resource use and CO_2 emissions before clicking on it. We therefore use survey responses as proxy for the level of engagement with the feedback email. Appendix Table A15 shows response rates by treatment group in the uploaders-only sample. The overall response rate among uploaders was 87% for the first email and 71% for the second email. While the share of respondents in the SER group was 8.4%p lower than in the DUAL group for the first email ($p = 0.203$) and 9.4%p higher for the second mail ($p = 0.308$), both differences are statistically insignificant. Furthermore, we find no evidence that uploaders in the DUAL group studied the reports more carefully than uploaders in SER group. Table A15 also compares estimation error across treatment group, defined as percent deviation of the water use estimate in the mini-survey from the exact

[33] An opposite effect is also conceivable, in which paying attention to one intervention decreases engagement with the other, for example due to cognitive capacity constraints (see, e.g., Altmann et al., 2022; Trachtmann, 2022) or lower perceived marginal benefits of information when subjects already receive real-time feedback. This would work against our complementarity argument.

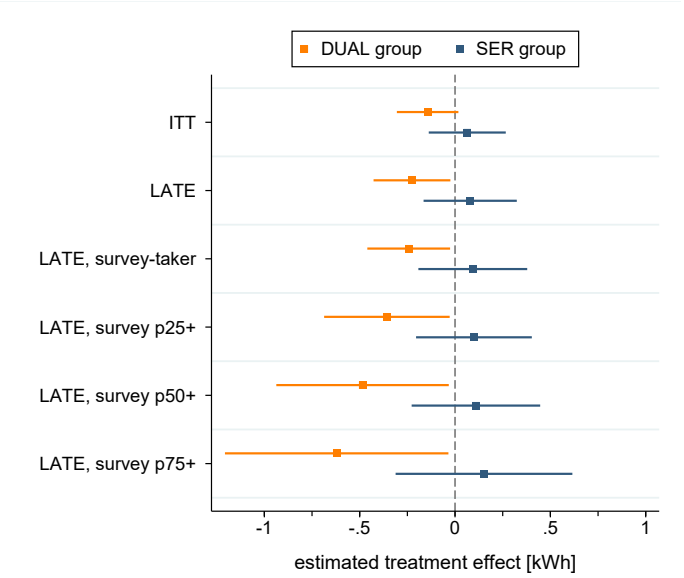


Figure 8: Effects for different levels of engagement with shower energy reports

Notes. The squares represent estimated regression coefficients for the effects of shower energy reports in intervention stage 2, where treatment engagement status is instrumented with treatment assignment (with the exception of ITT). Lines represent 90% confidence intervals. LATE, survey all includes all subjects who uploaded data and clicked on at least one mini-survey. The labels p25+/p25+/p75+ denote the groups of subjects whose estimate precision, defined as distance between estimated and measured water use per shower calculated for the shower energy reports, was above the 25th, 50th, or 75th percentile of all subjects, respectively. Responses from the two mini-surveys are combined by using the minimum estimate precision to define indicators.

number that we showed in the same personalized report that contained the survey link.[34] Smaller estimation errors are thus a direct indication of subjects paying closer attention while reading. Unsurprisingly, all treated groups outperformed the control group, but we observe no significant difference between uploaders in the SER and the DUAL group.

As a final plausibility check that differences between SER and DUAL are not driven by differential level of engagement with the shower energy reports, we look at whether subjects who studied the reports more closely also engaged more strongly in conservation actions. To do so, we create new treatment compliance indicators that also consider subject's level of engagement with the reports, to varying degrees of strictness. Specifically, we define an indicator for whether subjects uploaded data *and* clicked on the mini-survey in their report, and additional indicators for whether a subjects' estimation precision in the mini-surveys (as defined above) was above the 25th, 50th, or 75th percentile, respectively, of their treatment group. To avoid the endogeneity issue at hand, we instrument each of these treatment compliance indicators with the randomized assignment. Figure 8 plots results for the effect of shower energy reports in SER and DUAL group, respectively, when using these new set of indicators. The estimated conservation effect in the DUAL

[34] For the CON and RTF group, we calculate this using the number that we would have shown, were the subjects assigned to one of the shower energy report groups instead. Similarly, for non-uploaders, we use the ex post analogue of the same statistic, based on data was uploaded after the report or manually read out by our research team.

group increases monotonically with the strictness of our compliance definition, reaching more than 0.5 kWh per shower for the strictest 75th percentile indicator. In contrast, even the most studious subjects in the SER group did not reduce their energy use on average. Overall, it is therefore unlikely that our empirical results can be explained by differential level of engagement with the shower energy reports.

7.3. Other potential mechanisms

There are a number of alternative channels through which our interventions could affect conservation behavior. Hawthorne effects are one possibility, but recall that also subjects in the control group received a smart meter and emails reminding them to upload their data. Another potential channel may be that we accidentally killed the joy of showering. Reassuringly, our endline survey results suggest that the interventions had no effects on self-reported shower comfort, thus also alleviating concerns about unintended negative welfare effects (e.g., Damgaard and Gravert, 2018; Allcott and Kessler, 2019). Furthermore, we find no evidence for positive effects on general pro-environmental attitudes. If anything, we observe a decrease in self-perceived environmentalism in the treated groups compared to the control group, potentially due to feedback provision curbing the capacity for distorted self-image formation. See Appendix F and Table A17 for all results and a more detailed discussion of alternative mechanisms.

8. Discussion

In this paper, we argued that when multiple biases arising from different sources (e.g., imperfect information, limited attention) simultaneously prevent individuals from acting on their values and intentions, then combining interventions that each target a different source of bias can result in complementarity, meaning that each intervention becomes more effective when implemented in conjunction with the other(s) than in isolation. We first introduced a novel theoretical framework that explores the implications of multiple biases and defines conditions for complementary in behavioral interventions in such a setting. We then presented results from a three-month field experiment on resource conservation behavior in an energy- and water-intensive everyday activity (showering), in which we tested the effects of two types of interventions: shower energy reports, which provided information on resource use and carbon emissions via email, and real-time feedback, which made resource use in the shower immediately salient through a smart meter display. While only the latter induced a significant conservation effect when implemented in isolation, combining both interventions resulted in a striking complementarity that is in line with theoretical predictions. Specifically, it seems that knowledge gains about environmental impacts of water heating only translated into behavioral change when resource use was additionally made salient through real-time feedback.

8.1. Relevance of effect sizes

Although our interventions were targeted towards one specific resource-intensive activity, showering, the effect sizes are also quantitatively meaningful on the aggregate household level, which is all the more remarkable given that our subjects had no monetary incentives to conserve resources. In our study, real-time feedback in isolation lowered consumption by 0.4 kWh (6.3 liters) per shower; adding shower energy reports further lowered consumption by 0.23 kWh (3.8 liters). For comparison, total daily energy use for lighting in German households was less than 0.35 kWh per person on average at that time.[35] In his influential evaluation of the Opower home energy reports, which target *aggregate* electricity use in U.S. households, Allcott (2011) finds an average household-level conservation effect of 0.62 kWh per day.

For a simple cost-benefit calculation of the Amphiro smart shower meters, we refer to Tiefenbeck et al. (2018). For the shower energy reports, while it would not be credible to state a generally applicable estimate for the cost of intervention, we note that given we had already set up the technical infrastructure for real-time feedback in our study, the marginal costs of adding information on environmental impacts through emails were close to zero. Although we find no evidence that the email interventions were (cost-)effective in isolation, they produced significant additional conservation effect when combined with real-time feedback, so the bundled intervention should be the most cost-effective regime. This possibility was proposed theoretically in subsection 2.3.

8.2. Data limitations

Our data has a number of limitations. One issue is that the Amphiro devices had no global time counter, so our only source of information on timing comes from data uploads through a smartphone app. As most subjects tended to upload data in batches, some uncertainty remains about when exactly a shower took place, implying that we cannot easily control for date or time of day fixed effects and that there is some fuzziness around which observations took place before or after a shower energy report — to address this, we conducted a number of robustness checks. Moreover, a subset of participants could not upload any data due to technical issues with the Bluetooth connection, and thus could not receive any informative report. In principle, such problems that happen in early stages in the life cycle of new applications can be ironed out in the future.

Another limitation is that we cannot measure behavior outside of the shower. Hence, we cannot directly rule out, for example, whether subjects substitute part of their hygiene behavior to other facilities such as gym showers or wash basins. However, we find no evidence of extensive margin effects (i.e., on the number of observed showers)

[35] Source: German Federal Statistical Office (`https://www.destatis.de/EN/Themes/Society-Environment/Environment/Environmental-Economic-Accounting/private-households/Tables/energy-consumption-households.html`)

across experimental conditions, and a related study on water conversation in dorm showers finds no evidence of students moving between different communal shower facilities within the same building (Goette et al., 2021). Relatedly, we cannot account for potential spillover effects on other resource consumption activities, e.g., kitchen water usage or room heating. It is unclear whether this leads to an overstatement or an understatement of the overall impact of our interventions, and the general evidence for spillover effects of pro-environmental interventions is mixed (e.g., Tiefenbeck et al., 2013; Jessoe et al., 2021; Goetz et al., 2021; Sherif, 2021). Exploring the direction and magnitude of spillover effects thus constitutes an important avenue for future behavioral research.

8.3. Generalizability

Our study on hot water conservation in student dorm showers constitutes a very specific setting. First, students may generally not be representative of the general population. As our subjects also self-selected into participating in the study, they may on average be more intrinsically motivated to protect the environment, although note that about two-thirds of all dorm residents that we could reach through door-to-door recruitment participated in our study. Another noteworthy feature is that students in our sample did not have to pay utility bills and thus had no monetary incentives to conserve water and energy, which is unusual but not unheard of in other settings (e.g., office energy use).[36] The theoretical predictions for how these factors would affect the potential for complementarity are ambiguous. On the one hand, stronger marginal (monetary or non-monetary) incentives gives more leverage to alleviating informational and behavioural barriers; on the other hand, ceiling effects may limit the conservation potential if individuals already put in more effort in baseline. Our study was also conducted during winter in Germany, where demand for long and hot showers may have been higher due to cold weather. This could have limited conservation effects due to lower willingness to reduce warm water use, but also increased conservation potential due to a higher baseline. Another characteristic of our sample is that all subjects lived in single-person flats in relatively large and anonymous dorm buildings. This has advantages for the empirical study design, but limits the extent of information sharing and social influence that could be relevant in multi-person households. Interestingly, Tiefenbeck et al. (2018) found no difference in effects of real-time feedback between one- and two-person households.

Finally, whether and to which extent similar results would arise in other behavioral contexts is an open question. Our theoretical framework predicts that complementarities should become more likely if — following the Anna Karenina principle — multiple different mechanisms play a role in preventing behavioral change, including information frictions and behavioral biases like limited attention and self-control problems, but also

[36] Ito et al. (2018) find that monetary incentives lead to stronger and more persistent reductions in peak electricity use than to moral appeals. Also note that some surveys in Germany find that young people were less likely to behave sustainable in their daily lives than older generations (e.g., Ipsos, 2019).

more standard economic barriers such as lack of incentives or constraints on time, money, or technology. We suspect that such complexity of behavioral mechanisms is a pervasive feature of many social and economic domains of our lives, e.g., decisions affecting environmental, financial, or health outcomes. If true, the Anna Karenina effect we highlight in this study could imply the existence of numerous untapped opportunities for more targeted intervention designs and help organize empirical research on interaction effects of behavioral policy.[37] Obviously, our study cannot offer any definitive conclusion and should be viewed more as a proof of concept. More research is needed to understand how our findings would extrapolate to other settings and samples.

8.4. Implications for policy and research

Policy evaluation typically requires to test whether an intervention in isolation leads to the desired outcomes, to avoid that other interventions confound the effect. Similarly, behavioral researchers who wish to investigate specific determinants of behavior need to manipulate these determinants of interest while holding all other factors constant. However, our study highlights a particular generalizability challenge. Lack of observable impacts in response to an intervention (or manipulation) in isolation may be insufficient to rule out that it is not relevant or effective even within the same sample. For example, one may have concluded from the lack of effectiveness of our shower energy reports in isolation that improving knowledge about energy- and carbon-intensity of water heating does not matter in our context, but our findings suggest that knowledge gains may have been prevented from inducing observable behavioral change by residual biases like inattention or present bias.

The reason for this is that any singular evaluation of an intervention is inevitably confined to the particular choice environment it is introduced into, which is shaped by existing policies, institutions, norms, and individual circumstances. This environment itself can be malleable, so interventions that seem feeble at first glance may be able to unfold their full potential only once combined with complementary policies. We focused specifically on the role of multiple behavioral biases in decision-making creating potential for complementarities, because mitigating one specific bias (e.g., due to knowledge gaps) becomes more effective when also mitigating residual biases (e.g., due to inattention, self-control problems). This implies that policy designers should not focus only narrowly on one specific behavioral mechanism, but also attempt to identify other behavioral factors and how they might interact or interfere. For example, our results suggest that giving people tools that allow them to track their resource use may also make behavior more

[37] For example, Dupas and Robinson (2013) study financial behavior in a development context and find that simply providing a safe box for storing money is effective for encouraging higher savings, except for the subgroup of individuals with severe present bias, who need additional social commitment. Similarly, prompting deliberation about food choice to help resist temptations increases the effectiveness of healthy purchasing subsidies (Brownback et al., 2019). Cortes et al. (2023) find that text-message interventions on parenting practices work less well when in time periods where parents face high cognitive load.

sensitive to other policies such as informational and norm-based interventions; the same may also apply to conventional policies like price incentives (Jessoe and Rapson, 2014). An interesting approach for future research could be to first identify and elicit the existence and strength of different behavioral motives and biases at the individual level, and then implement and test tailored combinations of interventions in a second step — akin to personalized medical diagnoses and prescriptions.

The potential for complementarities creates a trade-off for policy makers with a binding budget constraint. They could either target more people with a single intervention or fewer people with a bundled intervention. We show in our theoretical framework that when complementarities between interventions are sufficiently strong, it can be preferable to implement a bundled approach at the cost of covering fewer households. As social scientists are beginning to pioneer the process from small-scale proof-of-concept studies to large-scale interventions (Banerjee et al., 2017), future research should therefore synchronously advance our knowledge on the interplay of different policy instruments.

References

ABRAHAMSE, W., L. STEG, C. VLEK, AND T. ROTHENGATTER (2005): "A Review of Intervention Studies Aimed at Household Energy Conservation," *Journal of Environmental Psychology*, 25, 273–291.

AGARWAL, S., X. FANG, L. GOETTE, S. SCHOEB, T. STAAKE, V. TIEFENBECK, AND D. WANG (2020): "The Role of Goals in Motivating Behavior: Evidence from a Large-Scale Field Experiment on Resource Conservation," *mimeo*.

ALLCOTT, H. (2011): "Social Norms and Energy Conservation," *Journal of Public Economics*, 95, 1082–1095.

——— (2016): "Paternalism and Energy Efficiency: An Overview," *Annual Review of Economics*, 8, 145–176.

ALLCOTT, H. AND J. B. KESSLER (2019): "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons," *American Economic Journal: Applied Economics*, 11, 236–276.

ALLCOTT, H. AND T. ROGERS (2014): "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, 104, 3003–3037.

ALTMANN, S., A. GRUNEWALD, AND J. RADBRUCH (2022): "Interventions and cognitive spillovers," *The Review of Economic Studies*, 89, 2293–2328.

ANDOR, M., A. GERSTER, J. PETERS, AND C. M. SCHMIDT (2020): "Social Norms and Energy Conservation Beyond the US," *Journal of Environmental Economics and Management*, 103, 102351.

ANDOR, M. A. AND K. M. FELS (2018): "Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects," *Ecological Economics*, 148, 178–210.

ASHRAF, N., B. K. JACK, AND E. KAMENICA (2013): "Information and Subsidies: Complements or Substitutes?" *Journal of Economic Behavior & Organization*, 88, 133–139.

ATTARI, S. Z. (2014): "Perceptions of Water Use," *Proceedings of the National Academy of Sciences of the United States of America*, 111, 5129–5134.

ATTARI, S. Z., M. L. DEKAY, C. I. DAVIDSON, AND W. BRUINE DE BRUIN (2010): "Public Perceptions of Energy Consumption and Savings," *Proceedings of the National Academy of Sciences of the United States of America*, 107, 16054–16059.

BANERJEE, A., R. BANERJI, J. BERRY, E. DUFLO, H. KANNAN, S. MUKERJI, M. SHOTLAND, AND M. WALTON (2017): "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, 31, 73–102.

BANERJEE, A., A. CHANDRASEKHAR, S. DALPATH, E. DUFLO, J. FLORETTA, M. JACKSON, H. KANNAN, F. LOZA, A. SANKAR, A. SCHRIMPF, AND M. SHRESTHA (2021): "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization," *Working Paper*.

BERNHEIM, B. D. AND D. TAUBINSKY (2018): "Behavioral public economics," *Handbook of behavioral economics: Applications and Foundations 1*, 1, 381–516.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2022): "Salience," *Annual Review of Economics*, 14, 521–544.

BRANDON, A., P. FERRARO, J. A. LIST, R. METCALFE, M. PRICE, AND F. RUNDHAMMER (2017): "Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments," *NBER Working Paper*, w23277.

BRANDON, A., J. A. LIST, R. D. METCALFE, M. K. PRICE, AND F. RUNDHAMMER (2019): "Testing for Crowd Out in Social Nudges: Evidence from a Natural Field Experiment in the Market for Electricity," *Proceedings of the National Academy of Sciences of the United States of America*, 116, 5293–5298.

BROWNBACK, A., A. IMAS, AND M. A. KUHN (2019): "Behavioral food subsidies," *The Review of Economics and Statistics*, 1–47.

BYRNE, D. P., L. GOETTE, L. A. MARTIN, A. JONES, A. MILES, S. SCHOB, T. STAAKE, AND V. TIEFENBECK (2022): "The Habit-Forming Effects of Feedback: Evidence from a Large-Scale Field Experiment," 70.

BYRNE, D. P., A. LA NAUZE, AND L. A. MARTIN (2018): "Tell Me Something I Don't Already Know: Informedness and the Impact of Information Programs," *Review of Economics and Statistics*, 100, 510–527.

CAMILLERI, A. R., R. P. LARRICK, S. HOSSAIN, AND D. PATINO-ECHEVERRI (2019): "Consumers Underestimate the Emissions Associated with Food but are Aided by Labels," *Nature Climate Change*, 9, 53–58.

CARLSSON, F., C. A. GRAVERT, V. KURZ, AND O. JOHANSSON-STENMAN (2021): "The Use of Green Nudges as an Environmental Policy Instrument," *Review of Environmental Economics and Policy*, 15, 216–237.

COE, D. T. AND D. J. SNOWER (1997): "Policy Complementarities: The Case for Fundamental Labor Market Reform," *IMF Staff Papers*, 44.

CORTES, K. E., H. FRICKE, S. LOEB, D. S. SONG, AND B. N. YORK (2023): "When behavioral barriers are too high or low–How timing matters for text-based parenting interventions," *Economics of Education Review*, 92, 102352.

DAMGAARD, M. T. AND C. GRAVERT (2018): "The hidden costs of nudging: Experimental evidence from reminders in fundraising," *Journal of Public Economics*, 157, 15–26.

DELMAS, M. A., M. FISCHLEIN, AND O. I. ASENSIO (2013): "Information Strategies and Energy Conservation Behavior: A Meta-analysis of Experimental Studies from 1975 to 2012," *Energy Policy*, 61, 729–739.

DENA (2016): "dena-Gebäudereport–Statistiken und Analysen zur Energieeffizienz im Gebäudebestand," .

DUFLO, E., P. DUPAS, AND M. KREMER (2015): "Education, HIV, and Early Fertility: Experimental Evidence from Kenya," *American Economic Review*, 105, 2757–2797.

DUPAS, P., E. HUILLERY, AND J. SEBAN (2018): "Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon," *Journal of Economic Behavior & Organization*, 145, 151–175.

DUPAS, P. AND J. ROBINSON (2013): "Why Don't the Poor Save More? Evidence from Health Savings Experiments," *American Economic Review*, 103, 1138–1171.

DADDA, G., Y. GAO, AND M. TAVONI (2020): "Making Energy Costs Salient Can Lead to Low-Efficiency Purchases," *E2e Working Paper 045*.

EUROPEAN ENVIRONMENT AGENCY (2021): "Water Resources Across EuropeConfronting Water Stress: An Updated Assessment," .

FANGHELLA, V., M. PLONER, AND M. TAVONI (2021): "Energy saving in a simulated environment: An online experiment of the interplay between nudges and financial incentives," *Journal of Behavioral and Experimental Economics*, 93, 101709.

FERRARO, P. J. AND M. K. PRICE (2013): "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment," *Review of Economics and Statistics*, 95, 64–73.

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

FREDERIKS, E. R., K. STENNER, AND E. V. HOBMAN (2015): "Household Energy Use: Applying Behavioural Economics to Understand Consumer Decision-Making and Behaviour," *Renewable and Sustainable Energy Reviews*, 41, 1385–1394.

GABAIX, X. (2017): "Behavioral Inattention," *NBER Working Papers 24096*.

GARDNER, G. T. AND P. C. STERN (2008): "The Short List: The Most Effective Actions U.S. Households Can Take to Curb Climate Change," *Environment and Behavior*, 50, 12–24.

GERSTER, A., M. ANDOR, AND L. GOETTE (2020): "Disaggregate Consumption Feedback and Energy Conservation," *CEPR Discussion Paper 14952*.

GIACCHERINI, M., D. H. HERBERICH, D. JIMENEZ-GOMEZ, J. A. LIST, G. PONTI, AND M. K. PRICE (2020): "Are Economics and Psychology Complements in Household Technology Diffusion? Evidence from a Natural Field Experiment," *Working Paper*.

GOETTE, L., H.-J. HAN, Z. H. LIM, ET AL. (2021): "The Dynamics of Goal Setting: Evidence From a Field Experiment on Resource Conservation," Tech. rep., University of Bonn and University of Mannheim, Germany.

GOETZ, A., H. MAYR, AND R. SCHUBERT (2021): "Beware of Side Effects? Spillover Evidence from a Hot Water Intervention," *Working Paper*.

HAHN, R., R. D. METCALFE, D. NOVGORODSKY, AND M. K. PRICE (2016): "The Behavioralist as Policy Designer: The Need to Test Multiple Treatment to Meet Multiple Targets," *NBER Working Paper 22886*.

HANNA, R., S. MULLAINATHAN, AND J. SCHWARTZSTEIN (2014): "Learning Through Noticing: Theory and Evidence from a Field Experiment," *Quarterly Journal of Economics*, 129, 1311–1353.

HOLLADAY, J. S., J. LARIVIERE, D. NOVGORODSKY, AND M. PRICE (2019): "Prices versus nudges: What matters for search versus purchase of energy investments?" *Journal of Public Economics*, 172, 151–173.

IMAI, T., D. D. PACE, P. SCHWARDMANN, AND J. J. VAN DER WEELE (2022): "Correcting Consumer Misperceptions About CO 2 Emissions," .

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467.

IPSOS (2019): "Ältere Leben Umweltbewusster Als Die Jugend - Aber Umweltverhalten Ändert Sich Bei Den Jungen Am Stärksten," Press release (28 august 2019).

ITO, K., T. IDA, AND M. TANAKA (2018): "Moral suasion and economic incentives: Field experimental evidence from energy demand," *American Economic Journal: Economic Policy*, 10, 240–67.

JAMISON, J. C., D. KARLAN, AND J. ZINMAN (2014): "Financial Education and Access to Savings Accounts: Complements or Substitutes? Evidence from Ugandan Youth Clubs," *NBER Working Paper 20135*.

JESSOE, K., G. E. LADE, F. LOGE, AND E. SPANG (2021): "Spillovers from Behavioral Interventions: Experimental Evidence from Water and Energy Use," *Journal of the Association of Environmental and Resource Economists*, 8, 315–346.

JESSOE, K. AND D. RAPSON (2014): "Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use," *American Economic Review*, 104, 1417–1438.

KARLIN, B., J. F. ZINGER, AND R. FORD (2015): "The Effects of Feedback on Energy Conservation: A Meta-analysis," *Psychological Science*, 141, 1205–1227.

KHANNA, T. M., G. BAIOCCHI, M. CALLAGHAN, F. CREUTZIG, H. GUIAS, N. R. HADDAWAY, L. HIRTH, A. JAVAID, N. KOCH, S. LAUKEMPER, ET AL. (2021): "A multicountry meta-analysis on the role of behavioural change in reducing energy consumption and CO2 emissions in residential buildings," *Nature Energy*, 6, 925–932.

KOLLMUSS, A. AND J. AGYEMAN (2002): "Mind the Gap: Why Do People Act Environmentally and What Are the Barriers to Pro-Environmental Behavior?" *Environmental Education Research*, 8, 239–260.

LIST, J. A., R. D. METCALFE, M. K. PRICE, AND F. RUNDHAMMER (2017): "Harnessing Policy Complementarities to Conserve Energy: Evidence from a Natural Field Experiment," *NBER Working Paper 23355*.

MBITI, I., K. MURALIDHARAN, M. ROMERO, Y. SCHIPPER, C. MANDA, AND R. RAJANI (2019): "Inputs, Incentives, And Complementarities In Education: Experimental Evidence From Tanzania," *Quarterly Journal of Economics*, 134, 1627–1673.

MURALIDHARAN, K., M. ROMERO, AND K. WÜTHRICH (2020): "Factorial designs, model selection, and (incorrect) inference in randomized experiments," *The Review of Economics and Statistics*, 1–44.

MYERS, E. AND M. SOUZA (2019): "Social Comparison Nudges Without Monetary Incentives: Evidence from Home Energy Reports," *E2e Working Paper 041*.

NOLAN, J. M., P. W. SCHULTZ, R. B. CIALDINI, N. J. GOLDSTEIN, AND V. GRISKEVICIUS (2008): "Normative Social Influence is Underdetected," *Personality and Social Psychology Bulletin*, 34, 913–923.

RUBIN, M. (2021): "When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing," *Synthese*, 199, 10969–11000.

SCHWARTZ, D. AND G. LOEWENSTEIN (2017): "The Chill of the Moment: Emotions and Proenvironmental Behavior," *Journal of Public Policy & Marketing*, 36, 255–268.

SHERIF, R. (2021): "Are Pro-environment Behaviours Substitutes or Complements? Evidence from the Field," *Max Planck Institute for Tax Law and Public Finance Working Paper 2021 – 03*.

TIEFENBECK, V. (2016): "On the Magnitude and Persistence of the Hawthorne Effect — Evidence from Four Field Studies," *4th European Conference on Behaviour and Energy Efficiency, Coimbra, Portugal*.

TIEFENBECK, V., L. GOETTE, K. DEGEN, V. TASIC, E. FLEISCH, R. LALIVE, AND T. STAAKE (2018): "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation," *Management Science*, 64, 1458–1476.

TIEFENBECK, V., T. STAAKE, K. ROTH, AND O. SACHS (2013): "For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign," *Energy Policy*, 57, 160–171.

TIEFENBECK, V., A. WOERNER, S. SCHOEB, E. FLEISCH, AND T. STAAKE (2019): "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives," *Nature Energy*, 4, 35–41.

TOLSTOY, L. (2003): *Anna Karenina*, (First published in Russian, 1873-1877; translation by Richard Pevear and Larissa Volokhonsky), London: Penguin Books.

TONKE, S. (2019): "Imperfect Knowledge, Information Provision and Behavior: Evidence from a Field Experiment to Encourage Resource Conservation," *Working Paper*.

TRACHTMANN, H. (2022): "Does promoting one behavior distract from others? Evidence from a field experiment," Tech. rep.

WICHMAN, C. J. (2017): "Information provision and consumer behavior: A natural experiment in billing frequency," *Journal of Public Economics*, 152, 13–33.

YOUNG, A. (2019): "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *Quarterly Journal of Economics*, 134, 557–598.

For Online Publication

Appendix A Supplementary figures and tables

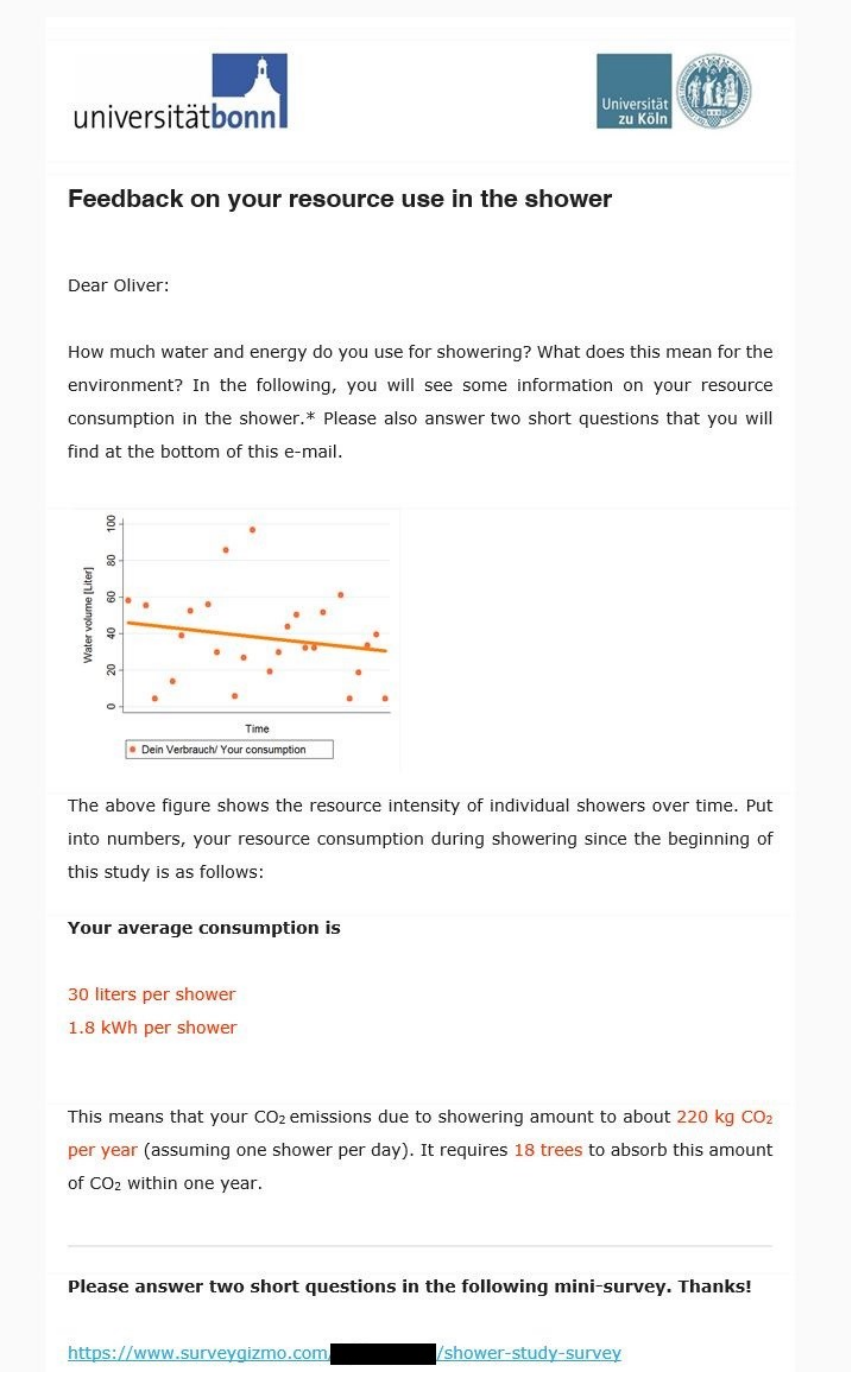


Figure A1: Screenshot of a typical shower energy report (for a fictitious person)

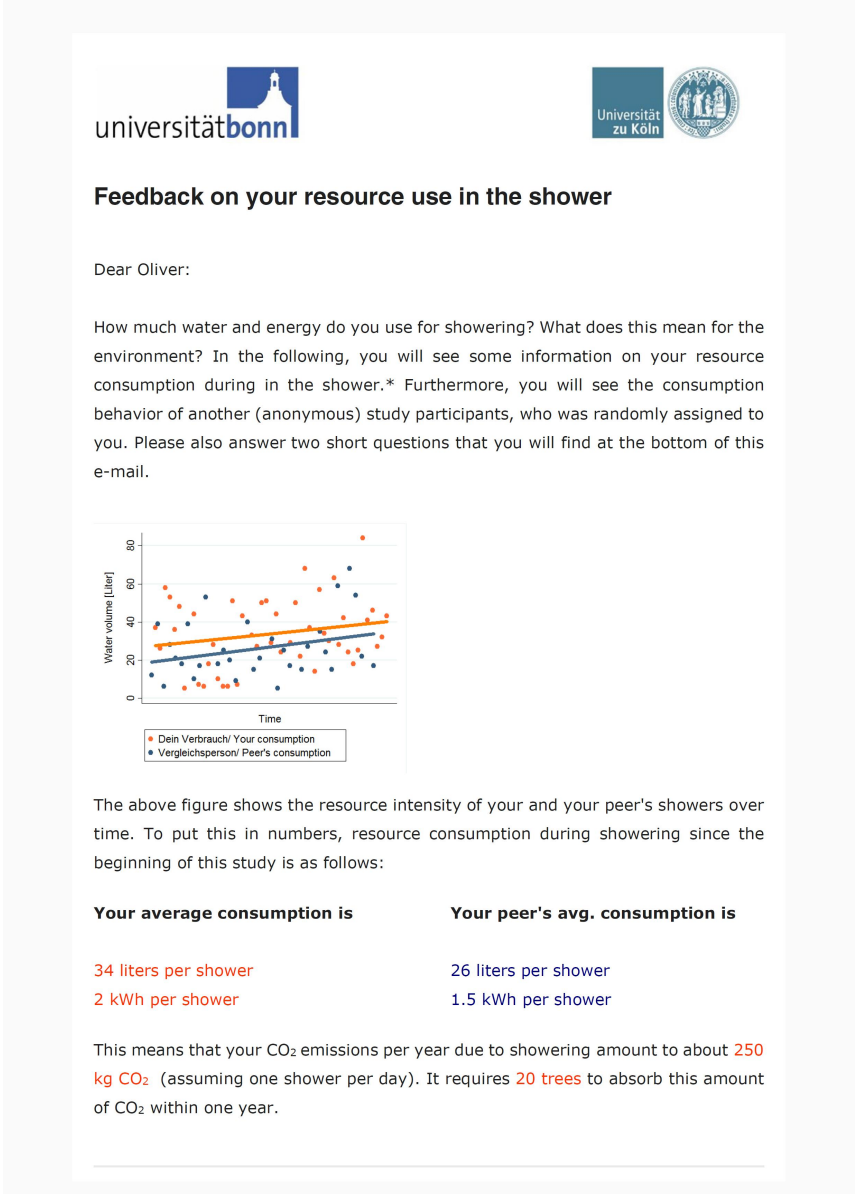


Figure A2: Screenshot of a shower energy report with peer comparison

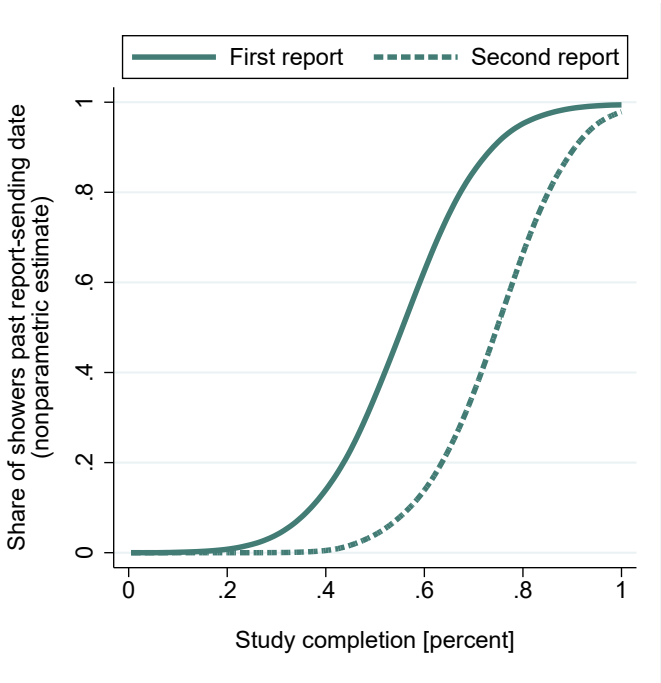


Figure A3: Empirical distribution of report timing

Table A1: Additional randomization checks

	Baseline survey responses				
	environmental attitude	shower comfort	1 if female	age in years	1 if international
SER group	-0.106	0.094	-0.046	0.757	-0.017
	(0.165)	(0.164)	(0.080)	(0.615)	(0.075)
RTF group	0.044	-0.164	-0.015	0.872	0.042
	(0.167)	(0.156)	(0.079)	(0.584)	(0.077)
DUAL group	0.154	0.115	0.117	0.540	0.032
	(0.161)	(0.149)	(0.075)	(0.583)	(0.075)
Constant	-0.041	-0.014	0.597	23.351	0.325
	(0.118)	(0.100)	(0.056)	(0.380)	(0.054)
Observations	307	306	318	307	318
R-squared	0.009	0.012	0.017	0.007	0.003
F-test: p-value	0.425	0.327	0.130	0.437	0.847

Robust standard errors in parentheses. The omitted category is the CON group.

Table A2: Balance checks for the TOT sample

	Volume [Liter]	Energy [kWh]	Duration [min]	Temperature [řC]	Flow rate [l/min]
SER group	-0.116	0.050	0.521	1.346	-0.420
	(3.836)	(0.242)	(0.564)	(0.668)	(0.338)
RTF group	-2.366	-0.191	0.149	0.199	-0.154
	(3.533)	(0.220)	(0.537)	(0.632)	(0.384)
DUAL group	-0.671	-0.030	-0.110	0.583	0.101
	(4.154)	(0.257)	(0.579)	(0.613)	(0.421)
Constant	38.386	2.234	6.574	35.428	5.941
	(2.679)	(0.173)	(0.379)	(0.481)	(0.245)
Observations	260	260	260	260	260
R-squared	0.002	0.005	0.005	0.020	0.007
F-test: p-value	0.895	0.685	0.731	0.176	0.522

Robust standard errors in parentheses. The omitted category is the CON group.

Table A3: Additional balance checks for the TOT sample

	Baseline survey responses				
	environmental attitude	shower comfort	1 if female	age in years	1 if international
SER group	-0.135	0.110	-0.080	0.664	0.022
	(0.182)	(0.184)	(0.088)	(0.668)	(0.083)
RTF group	0.040	-0.191	-0.030	0.925	0.073
	(0.169)	(0.159)	(0.082)	(0.605)	(0.079)
DUAL group	0.129	-0.030	0.122	0.256	-0.060
	(0.181)	(0.175)	(0.084)	(0.658)	(0.079)
Constant	-0.033	-0.011	0.597	23.423	0.306
	(0.123)	(0.102)	(0.058)	(0.392)	(0.055)
Observations	257	257	261	257	261
R-squared	0.008	0.011	0.020	0.010	0.010
F-test: p-value	0.561	0.447	0.123	0.451	0.430

Robust standard errors in parentheses. The omitted category is the CON group.

Table A4: Effect of real-time feedback and ITT estimates

	only RTF & CON		*Intention-to-treat (ITT)*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
Intervention	0.283***	4.453***	0.200**	2.961*
	(0.104)	(1.597)	(0.099)	(1.530)
Intervention × RTF/DUAL	-0.397***	-6.346***	-0.368***	-5.514***
	(0.125)	(1.926)	(0.122)	(1.932)
Intervention × SER			-0.001	0.450
			(0.133)	(2.059)
Intervention × DUAL			0.121	2.366
			(0.108)	(1.753)
IN stage 2			0.152*	2.756**
			(0.090)	(1.360)
IN stage 2 × RTF/DUAL			-0.054	-1.550
			(0.113)	(1.806)
IN stage 2 × SER			0.064	0.464
			(0.120)	(1.883)
IN stage 2 × DUAL			-0.180*	-3.192*
			(0.098)	(1.710)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Clusters	156	156	318	318
Observations	8446	8446	17942	17942
R^2	0.379	0.375	0.403	0.404

Columns (1) and (2) only include individuals in the RTF or CON group. Standard errors in parentheses are clustered at the individual level. Permutation-based inference for the main coefficients of interest is depicted in Appendix Figure **??**.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Treatment on the treated (TOT) estimates with donut holes

	Uploaders-only		*LATE*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
Intervention	0.199*	2.945	0.188*	2.821*
	(0.118)	(1.806)	(0.109)	(1.662)
Intervention × RTF/DUAL	-0.400***	-5.983***	-0.381***	-5.773***
	(0.139)	(2.190)	(0.130)	(2.045)
Intervention × SER	0.011	0.591	0.005	0.500
	(0.158)	(2.458)	(0.138)	(2.139)
Intervention × DUAL	0.021	0.421	0.094	1.878
	(0.112)	(1.842)	(0.109)	(1.754)
IN stage 2	0.129	2.431	0.152	2.732*
	(0.098)	(1.492)	(0.095)	(1.432)
IN stage 2 × RTF/DUAL	-0.035	-1.367	-0.059	-1.595
	(0.121)	(1.967)	(0.117)	(1.889)
IN stage 2 × SER	0.115	1.220	0.106	1.064
	(0.138)	(2.167)	(0.154)	(2.403)
IN stage 2 × DUAL	-0.227**	-3.848**	-0.223*	-3.949*
	(0.109)	(1.926)	(0.121)	(2.135)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
p-value: $\gamma_2 = \beta_1$	0.015	0.030	0.025	0.043
p-value: $\gamma_2 = \gamma_3$	0.052	0.082	0.094	0.120
Observations	13465	13465	15584	15584
R^2	0.424	0.425	0.005	0.005

This table excludes shower observations with high uncertainty regarding whether they occurred before or after the first shower energy report. Specifically, we calculate post-report probabilities based on a binomial distribution, assuming that shower frequency remains constant, and only include observations with probability of either below 10% or above 90%. In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported R^2 in columns (3) and (4) is the within R^2. Standard errors in parentheses are clustered at the individual level. Permutation-based inference procedures are presented in Figure A4. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A6: TOT estimates with alternative timing definition for non-uploaders

	Uploaders-only		*LATE*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
Intervention	0.204*	3.004*	0.194*	2.868*
	(0.108)	(1.661)	(0.100)	(1.530)
Intervention × RTF/DUAL	-0.391***	-5.796***	-0.371***	-5.564***
	(0.130)	(2.062)	(0.121)	(1.924)
Intervention × SER	0.016	0.606	0.013	0.642
	(0.151)	(2.359)	(0.132)	(2.049)
Intervention × DUAL	0.038	0.621	0.111	2.175
	(0.110)	(1.799)	(0.105)	(1.704)
IN stage 2	0.114	2.253*	0.160*	2.848**
	(0.084)	(1.315)	(0.090)	(1.358)
IN stage 2 × DUAL	-0.239**	-3.980**	-0.234*	-4.115*
	(0.108)	(1.893)	(0.128)	(2.232)
IN stage 2 × SER	0.120	1.274	0.053	0.137
	(0.124)	(1.990)	(0.153)	(2.397)
IN stage 2 × RTF/DUAL	-0.018	-1.153	-0.047	-1.427
	(0.108)	(1.786)	(0.112)	(1.787)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
p-value: $\gamma_2 = \beta_1$	0.007	0.019	0.039	0.076
p-value: $\gamma_2 = \gamma_3$	0.030	0.057	0.152	0.195
Observations	14712	14712	17942	17942
R^2	0.412	0.415	0.004	0.004

This table uses an alternative timing definition in which *INstage*2 is coded as 1 for non-uploaders at the median study completion value of uploaders at the first report date. In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported R^2 in columns (3) and (4) is the within R^2. Standard errors in parentheses are clustered at the individual level. Permutation-based inference procedures are presented in Figure A4. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A7: TOT estimates using date of sending to define intervention stage 2

	Uploaders-only		*LATE*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
Intervention	0.186*	2.731*	0.184*	2.704*
	(0.104)	(1.608)	(0.096)	(1.480)
Intervention × RTF/DUAL	-0.381***	-5.685***	-0.366***	-5.506***
	(0.127)	(2.031)	(0.119)	(1.896)
Intervention × HER	0.021	0.698	0.008	0.586
	(0.151)	(2.365)	(0.131)	(2.042)
Intervention × DUAL	0.033	0.578	0.109	2.177
	(0.111)	(1.822)	(0.106)	(1.726)
IN stage 2	0.144*	2.693**	0.175*	3.096**
	(0.086)	(1.328)	(0.094)	(1.392)
IN stage 2 × RTF/DUAL	-0.034	-1.319	-0.055	-1.493
	(0.111)	(1.824)	(0.116)	(1.839)
IN stage 2 × HER	0.104	1.004	0.063	0.278
	(0.127)	(2.035)	(0.153)	(2.387)
IN stage 2 × DUAL	-0.225**	-3.803**	-0.220*	-3.973*
	(0.105)	(1.843)	(0.120)	(2.105)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
p-value: $\gamma_2 = \beta_1$	0.009	0.024	0.033	0.066
p-value: $\gamma_2 = \gamma_3$	0.047	0.081	0.146	0.183
Clusters	261	261	318	318
Observations	14712	14712	17942	17942
R^2	0.413	0.416	0.004	0.004

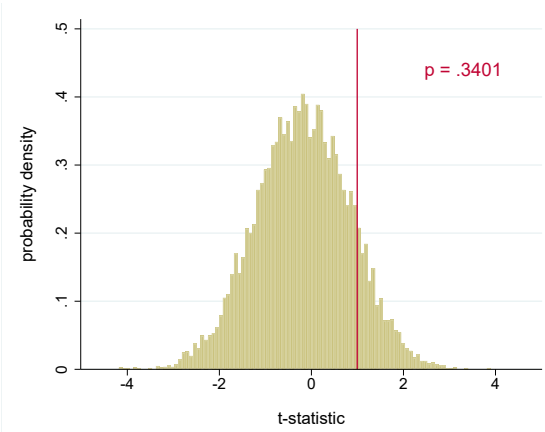
This table defines intervention stage 2 using only the date and time that shower energy reports (and placebo emails) were sent out. In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported R^2 in columns (3) and (4) is the within R^2. Standard errors in parentheses are clustered at the individual level. Permutation-based inference procedures are presented in Figure A4. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A8: Margins of behavioral adjustment

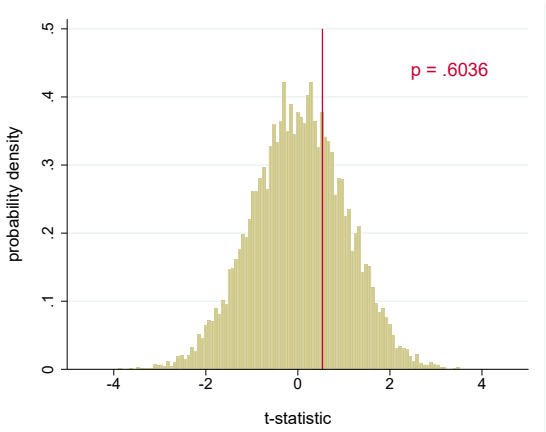
	Duration in seconds			Temperature in °C			Flow rate in liter/min		
	(1) ITT	(2) Uploaders	(3) LATE	(4) ITT	(5) Uploaders	(6) LATE	(7) ITT	(8) Uploaders	(9) LATE
Intervention	11.69	11.16	11.69	0.02	0.05	0.02	0.27**	0.26**	0.27**
	(10.67)	(11.32)	(10.67)	(0.33)	(0.36)	(0.33)	(0.11)	(0.12)	(0.11)
Intervention × RTF/DUAL	-41.63**	-41.39**	-41.63**	-0.71	-0.83*	-0.71	-0.12	-0.13	-0.12
	(16.11)	(16.88)	(16.11)	(0.45)	(0.47)	(0.45)	(0.16)	(0.17)	(0.16)
Intervention × SER	10.68	9.62	9.06	-0.51	-0.59	-0.50	-0.11	-0.15	-0.09
	(16.04)	(17.21)	(16.03)	(0.40)	(0.45)	(0.40)	(0.15)	(0.18)	(0.15)
Intervention × DUAL	7.11	7.77	6.98	-0.25	0.15	-0.27	-0.05	-0.25	-0.07
	(16.73)	(17.13)	(16.41)	(0.39)	(0.41)	(0.38)	(0.17)	(0.18)	(0.17)
IN stage 2	20.67	9.20	20.67	0.38	0.43	0.38	0.25**	0.23*	0.25**
	(16.54)	(9.33)	(16.54)	(0.28)	(0.30)	(0.28)	(0.11)	(0.12)	(0.11)
IN stage 2 × RTF/DUAL	-28.91	-17.53	-28.91	0.25	0.26	0.25	0.12	0.13	0.12
	(18.54)	(12.78)	(18.54)	(0.37)	(0.39)	(0.37)	(0.17)	(0.18)	(0.17)
IN stage 2 × SER	-28.54	-10.81	-35.28	0.12	0.17	0.15	0.23	0.17	0.28
	(20.39)	(16.23)	(25.49)	(0.34)	(0.37)	(0.43)	(0.17)	(0.17)	(0.21)
IN stage 2 × DUAL	-1.85	0.79	-2.33	-0.19	-0.37	-0.24	-0.29	-0.34	-0.36
	(12.38)	(13.80)	(15.58)	(0.35)	(0.33)	(0.43)	(0.21)	(0.23)	(0.26)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Observations	17942	14712	17942	17942	14712	17942	17942	14712	17942
R^2	0.383	0.361	0.001	0.309	0.322	0.003	0.751	0.763	0.015

For LATE specifications the within-R^2 is reported. Standard errors in parentheses clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

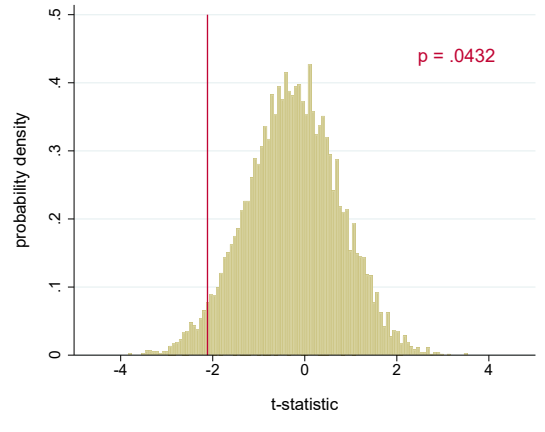
(a) Col 1: IN stage 2 × SER



(b) Col 3: IN stage 2 × SER



(c) Col 1: IN stage 2 × DUAL



(d) Col 3: IN stage 2 × DUAL

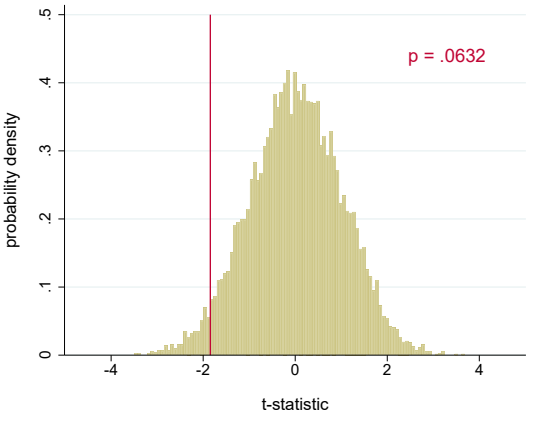
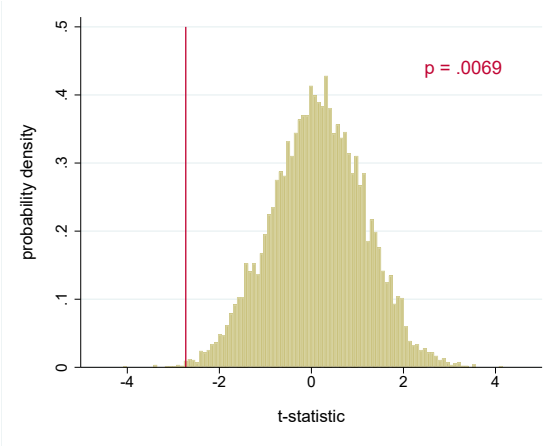


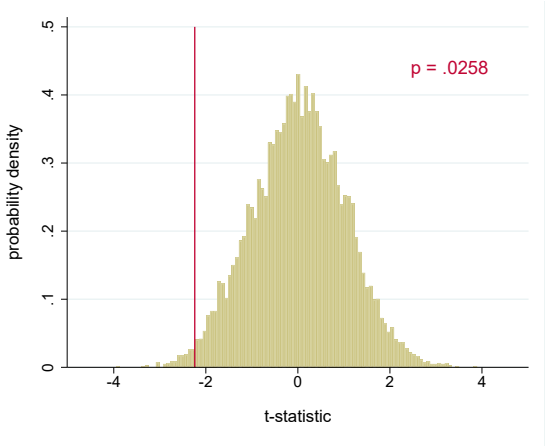
Figure A4: Randomization inference for coefficients of interest in Table 3

Notes. Distribution of estimated t-statistics based on 10,000 permutation samples. For each permutation, treatment assignment into CON, SER, RTF, or DUAL was randomly relabeled, holding constant the actual number of individuals in each treatment group. The red vertical line represents the t-value for the true treatment labels. Permutation-based p-values are shown in the top right corner.

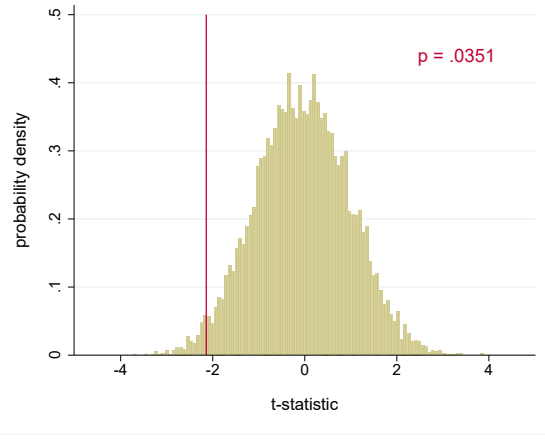
(a) Col 1: t-test for $H_0 : \gamma_2 = \beta_1$



(b) Col 3: t-test for $H_0 : \gamma_2 = \beta_1$



(c) Col 1: t-test for $H_0 : \gamma_2 = \gamma_3$



(d) Col 3: t-test for $H_0 : \gamma_2 = \gamma_3$

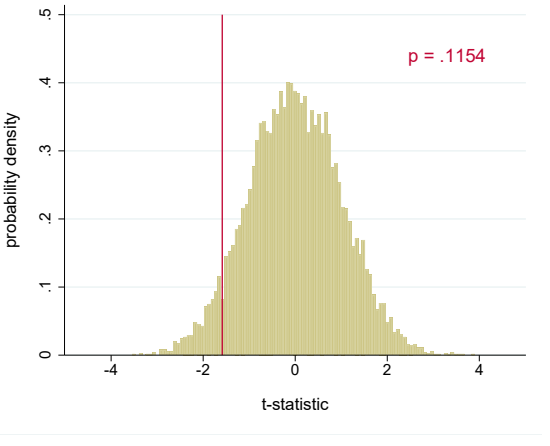


Figure A5: Randomization inference for additional hypothesis tests in Table 3

Notes. Distribution of estimated t-statistics based on 10,000 permutation samples. For each permutation, treatment assignment into CON, SER, RTF, or DUAL was randomly relabeled, holding constant the actual number of individuals in each treatment group. The red vertical line represents the t-value for the true treatment labels. Permutation-based p-values are shown in the top right corner.

Table A9: Treatment effect dynamics: all coefficients

	$Z_i = \mathbb{I}\{post\ 2nd\ report\}$		$Z_i = \#\ weeks\ after\ 1st\ report$	
	(1) Uploaders	(2) LATE	(3) Uploaders	(4) LATE
Intervention	0.208*	0.201**	0.206*	0.199**
	(0.108)	(0.099)	(0.108)	(0.099)
Intervention × RTF/DUAL	-0.388***	-0.368***	-0.387***	-0.367***
	(0.130)	(0.122)	(0.131)	(0.122)
Intervention × SER	0.012	0.003	0.014	0.005
	(0.151)	(0.132)	(0.151)	(0.132)
Intervention × DUAL	0.036	0.114	0.040	0.116
	(0.111)	(0.107)	(0.112)	(0.108)
IN stage 2	0.091	0.135	0.016	0.066
	(0.095)	(0.103)	(0.108)	(0.115)
IN stage 2 × RTF/DUAL	-0.038	-0.059	0.037	0.016
	(0.120)	(0.127)	(0.149)	(0.154)
IN stage 2 × SER	0.149	0.102	0.210	0.163
	(0.135)	(0.156)	(0.151)	(0.178)
IN stage 2 × DUAL	-0.086	-0.068	0.009	0.049
	(0.109)	(0.120)	(0.154)	(0.169)
IN stage 2 $\times Z\_i$	0.032	0.029	0.032	0.029
	(0.090)	(0.087)	(0.023)	(0.022)
IN stage 2 × RTF/DUAL $\times Z\_i$	0.025	0.009	-0.021	-0.024
	(0.128)	(0.125)	(0.035)	(0.034)
IN stage 2 × SER $\times Z\_i$	-0.043	-0.040	-0.030	-0.028
	(0.123)	(0.133)	(0.036)	(0.041)
IN stage 2 × DUAL $\times Z\_i$	-0.245	-0.273	-0.082	-0.095
	(0.201)	(0.207)	(0.055)	(0.058)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Observations	14712	17942	14712	17942
Clusters	261	318	261	318
R^2	0.413	0.004	0.413	0.005

Standard errors in parentheses are clustered at the individual level. In columns (1) and (2), we exclude all non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem. In columns (3) and (4), we use treatment assignment to SER and DUAL, respectively, interacted with the IN stage 2 indicator as instrument for receiving informative shower energy reports. The reported R^2 in Columns (3) and (4) is the within R^2.

$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Table A10: Treatment effect dynamics with trends and jumps

	(1) Uploaders	(2) LATE
Intervention	0.202*	0.195*
	(0.108)	(0.100)
Intervention × RTF/DUAL	-0.382***	-0.363***
	(0.131)	(0.122)
Intervention × HER	0.013	0.007
	(0.151)	(0.132)
Intervention × DUAL	0.039	0.116
	(0.112)	(0.108)
IN stage 2	0.008	0.039
	(0.100)	(0.109)
IN stage 2 × RTF/DUAL	0.051	0.046
	(0.156)	(0.161)
IN stage 2 × HER	0.111	0.073
	(0.149)	(0.188)
IN stage 2 × DUAL	0.032	0.074
	(0.181)	(0.192)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$	-0.154	-0.138
	(0.143)	(0.139)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × RTF/DUAL	0.221	0.200
	(0.183)	(0.180)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × SER	0.120	0.134
	(0.192)	(0.215)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × DUAL	-0.067	-0.051
	(0.229)	(0.235)
IN stage 2 × *Time trend*	0.066	0.076
	(0.057)	(0.056)
IN stage 2 × *Time trend* × RTF/DUAL	-0.070	-0.083
	(0.076)	(0.075)
IN stage 2 × *Time trend* × SER	0.033	0.029
	(0.086)	(0.105)
IN stage 2 × *Time trend* × DUAL	-0.089	-0.107
	(0.088)	(0.090)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × *Time trend*	0.003	-0.022
	(0.078)	(0.076)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × *Time trend* × RTF/DUAL	-0.003	0.020
	(0.097)	(0.095)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × *Time trend* × SER	-0.135	-0.137
	(0.137)	(0.153)
IN stage 2 × $\mathbb{I}\{$*post 2nd report*$\}$ × *Time trend* × DUAL	0.030	0.032
	(0.095)	(0.103)
Individual fixed effects	*yes*	*yes*
Observations	14712	17942
R^2	0.413	0.005

For LATE specifications the within-R^2 is reported. Standard errors in parentheses clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A11: Treatment effect heterogeneity by baseline consumption

	(1) continuous	(2) $\mathbb{I}\{> median\}$
Intervention × RTF/DUAL	-0.398***	-0.236**
	(0.124)	(0.096)
Intervention × RTF/DUAL × Baseline energy use	-0.183	-0.286
	(0.123)	(0.261)
IN stage 2 × RTF/DUAL	-0.019	0.151
	(0.110)	(0.103)
IN stage 2 × RTF/DUAL × Baseline energy use	-0.063	-0.344
	(0.102)	(0.218)
IN stage 2 × SER	0.129	0.193*
	(0.128)	(0.105)
IN stage 2 × SER × Baseline energy use	0.058	-0.149
	(0.129)	(0.248)
IN stage 2 × DUAL	-0.252**	-0.154
	(0.105)	(0.094)
IN stage 2 × DUAL × Baseline energy use	-0.165	-0.195
	(0.115)	(0.223)
Other treatment indicators	*yes*	*yes*
Individual fixed effects	*yes*	*yes*
Observations	14675	14675
R^2	0.414	0.413

The full table with all coefficients can be found in Appendix Table A12. All non-uploaders in SER and DUAL as well as all non-uploaders in RTF and CON who did not report a technical problem are excluded. Baseline energy use is demeaned, so main effects represent TEs at the sample mean. Standard errors in parentheses are clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A12: Treatment effect heterogeneity: all coefficients

	X_i : *baseline energy use*		X_i : *envir. attitude*	
	(1) linear	(2) median$^+$	(3) linear	(4) median$^+$
Intervention	0.206**	0.267***	0.208*	0.295
	(0.101)	(0.071)	(0.109)	(0.196)
Intervention × RTF/DUAL	-0.398***	-0.236**	-0.394***	-0.354*
	(0.124)	(0.096)	(0.131)	(0.214)
Intervention × SER	0.003	-0.085	-0.013	0.007
	(0.143)	(0.117)	(0.146)	(0.250)
Intervention × DUAL	0.086	0.013	0.046	-0.068
	(0.114)	(0.087)	(0.112)	(0.153)
IN stage 2	0.110	0.009	0.124	0.059
	(0.086)	(0.064)	(0.083)	(0.155)
IN stage 2 × RTF/DUAL	-0.019	0.151	-0.036	0.015
	(0.110)	(0.103)	(0.109)	(0.180)
IN stage 2 × SER	0.129	0.193*	0.110	0.271
	(0.128)	(0.105)	(0.122)	(0.202)
IN stage 2 × DUAL	-0.252**	-0.154	-0.237**	-0.366*
	(0.105)	(0.094)	(0.116)	(0.188)
Intervention $\times X_i$	0.018	-0.123	0.004	-0.183
	(0.108)	(0.222)	(0.119)	(0.211)
Intervention × RTF/DUAL $\times X_i$	-0.183	-0.286	-0.176	-0.113
	(0.123)	(0.261)	(0.135)	(0.261)
Intervention × SER $\times X_i$	0.055	0.189	-0.144	-0.030
	(0.138)	(0.298)	(0.158)	(0.285)
Intervention × DUAL $\times X_i$	0.110	0.054	0.088	0.265
	(0.129)	(0.216)	(0.101)	(0.232)
IN stage 2 $\times X_i$	-0.003	0.215	-0.034	0.138
	(0.083)	(0.170)	(0.100)	(0.168)
IN stage 2 × RTF/DUAL $\times X_i$	-0.063	-0.344	0.030	-0.103
	(0.102)	(0.218)	(0.116)	(0.221)
IN stage 2 × SER $\times X_i$	0.058	-0.149	0.040	-0.399*
	(0.129)	(0.248)	(0.129)	(0.233)
IN stage 2 × DUAL $\times X_i$	-0.165	-0.195	0.053	0.229
	(0.115)	(0.223)	(0.093)	(0.229)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Observations	14675	14675	14501	14501
Clusters	260	260	257	257
R^2	0.414	0.413	0.414	0.415

Standard errors in parentheses are clustered at the individual level. The coefficients are obtained using the within estimator. All non-uploaders in SER and DUAL, as well as all non-uploaders in RTF and CON who did not report a technical problem, are excluded. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A13: Estimated vs actual water use per shower

	before study	after study	
		ITT	TOT
Actual volume	0.271	0.199	0.215
	(0.263)	(0.139)	(0.147)
Actual volume × RTF	0.025	0.752***	0.838***
	(0.376)	(0.198)	(0.182)
Actual volume × SER	-0.465	0.273	0.363**
	(0.292)	(0.178)	(0.175)
Actual volume × DUAL	-0.074	0.500***	0.497**
	(0.299)	(0.181)	(0.237)
RTF group	-0.131	2.348	3.855
	(6.777)	(3.194)	(3.120)
SER group	-7.001	-4.575	-5.113*
	(5.813)	(3.105)	(2.948)
DUAL group	-5.182	2.039	2.050
	(5.851)	(3.118)	(3.939)
Constant	43.436***	39.556***	39.680***
	(4.590)	(2.347)	(2.449)
Observations	267	299	253
R^2	0.030	0.378	0.438

For column (1), actual volumes are approximated by the average baseline water usage per shower. For columns (2) and (3), actual volumes are approximated by the average water usage in the final third of observations. Actual volumes are recentered around 40 liters. Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A14: Estimated versus actual water use: relative estimation error

	before study	after study	
		ITT	TOT
RTF group	0.075	-0.271***	-0.282***
	(0.201)	(0.076)	(0.079)
SER group	0.008	-0.163**	-0.268***
	(0.175)	(0.082)	(0.075)
DUAL group	-0.055	-0.185**	-0.228***
	(0.178)	(0.087)	(0.079)
Constant	0.927***	0.569***	0.573***
	(0.136)	(0.064)	(0.067)
Observations	302	299	253
R^2	0.002	0.042	0.081

For column (1), actual volumes are approximated by the average baseline water usage per shower. For columns (2) and (3), actual volumes are approximated by the average water usage per shower in the final third of observations. Actual volumes are recentered around 40 liters. Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A15: Responses to mini-surveys attached to reports and placebo emails

	Survey response rate			Estimation error [%p]	
	(1) report 1	(2) report 2	(3) any report	(4) report 1	(5) report 2
RTF group	-1.05	0.53	-2.48	-43.42***	-31.73***
	(5.35)	(6.57)	(4.90)	(10.17)	(8.93)
SER group	-7.85	-16.76**	-7.18	-43.04***	-45.96***
	(6.39)	(7.91)	(5.81)	(13.53)	(8.22)
DUAL group	0.58	-26.17***	-0.44	-38.98***	-40.43***
	(5.54)	(8.14)	(5.00)	(11.60)	(8.50)
Constant	88.89***	80.56***	91.67***	71.18***	56.26***
	(3.73)	(4.70)	(3.28)	(9.75)	(7.88)
p-value for SER = DUAL	0.203	0.308	0.270	0.719	0.166
Observations	261	261	261	224	183
R^2	0.009	0.061	0.008	0.104	0.182

Columns 1 to 3 compares response rates to mini-surveys attached to the shower energy reports (or placebo emails for RTF and CON). Columns 4 and 5 compares average estimation errors relative to the measured water usage calculated for the reports, excluding extreme outliers who overestimate by a factor of more than 5. For non-uploaders, we use ex post calculations of the relevant water usage statistics instead. Non-uploaders in SER and DUAL as well as non-uploaders without technical problems in CON and RTF are excluded. Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix B Randomization protocol

At the beginning of the study, we randomly assigned subjects into groups that receive or do not receive real-time feedback. Each smart meter was programmed as either treatment or control device. Treatment device started displaying real-time feedback from the eleventh shower onwards, whereas control devices only ever showed the current water temperature. When distributing the smart meters to subjects, we alternated between treatment and control devices after each apartment. Thus, treatment and control devices are by construction balanced within dorms.

We assigned subjects into groups with or without shower energy report shortly before we intended to send out the reports. We used the data that subjects uploaded through the smartphone app to rank them from lowest to highest average water use per shower, split by whether they receive real-time feedback or not. Then, we formed pairs between subjects adjacent to each other in rank and assigned shower energy reports to only one member of a pair based on a virtual coin flip. This ensures that the distribution of resource consumption levels remain balanced across experimental conditions. Subjects who had not uploaded any data at that point in time were assigned to a group randomly without prior ranking.

The second shower energy report further contained a social comparison component with a random and anonymous peer. This peer was assigned to subjects in the following way: (1) we used uploaded data prior to the second report to rank subjects again by their average water use per shower; (2) we then selected three potential peers for each subject, a subject who was somewhat above him/her in rank, a subject who was somewhat below him/her in rank, and a directly adjacent subject; (3) we then chose one of these three candidates randomly with equal probabilities; (4) subjects who had not uploaded any data received a random peer from the pool of subjects who had uploaded data. This procedure ensured that the direction of peer comparison was orthogonal to subjects' resource use level.

Appendix C Data cleaning procedures

A number of data cleaning steps are performed before running the empirical analyses. In principle, we have access to the smart meter data from two sources: (1) uploads by subjects themselves using the smartphone app, and (2) the data that we read out manually after retrieving the devices. For the large majority of devices, the two sources gave us identical data. In the cases where it differed, we always opted to use the information we read out manually.

We drop the very first data point of each participant, as they usually started with a test run to check if the device was working. Following Tiefenbeck et al. (2018), we further drop any water extraction with volume below 4.5 liters (in total 2,942 extractions), as these are unlikely to be actual showers but rather minor extractions for other purposes

such as cleaning. We further remove 37 extreme outlier points, defined as energy use and water use for that shower being more than 4.5 times the subject-specific interquartile range away from the closest quartile. We are particularly strict in only excluding the most unplausible data points here. Conventionally, 1.5 or 3 times the interquartile range (IQR) are used as criterion for outliers. For a normal distribution, 4.5 times the IQR away from the nearest quartile corresponds to 6.745 standard deviation away from the mean.

We further exclude 1 device with erratic data, as evidenced by huge intra-device variance (the largest for all devices) and some outrageous data points with water volumes of up to above 500 liters for a single shower. In 8 cases, the device's temperature sensor broke at some point, and we impute missing information with the average temperature of showers taken while the sensor was still intact. For some devices, we detected an error through which decimal places of the flow rate are shifted such that the stored number is actually ten times the actual flow rate. We corrected these manually for showers with flow rates that are about ten times the flow rate of other showers stored on the device.

Appendix D Timing of showers

As the smart meter itself has no global time counter and only stores the chronological order of water extractions, we make use of smartphone app information to put a time stamp on each observation. In particular, we need to determine whether a shower took place before or after we sent out the shower energy reports, so whether it is in intervention stage 2. The app provides us with information on the date and time of each data upload by subjects. This allows us construct time windows in which a shower observation has plausibly happened. Firstly, a shower must have been taken by the time data was uploaded via the app, so this gives us the upper bound. Secondly, it must have been taken place after the previous data upload, because otherwise it would have been uploaded by then; this gives us the lower bound. To be able to determine the timing relatively reliably around the crucial time period, in which we sent out shower energy reports, we sent several upload reminders to all participants. Whenever it was not unambiguously clear, which shower was the first that took place after a shower energy report, we assigned the switching point implied by constant shower frequency. For example, if one upload was 1 day before the shower energy report and the next upload 1 day after, and there were 2 showers in the window, we assumed that the first shower was before and the second shower after the report.

A complication arising from non-uploaders is that we do not know the timing of showers by these participants, because the shower meter itself only stores the order of showers but not the time and date. We can only infer the earliest and latest possible date of each shower based on when it was uploaded to the smartphone app. Therefore, whenever we want to include non-uploaders in our analyses, we need to impute the timing of showers in one way or another, in particular whether it took place before or after a shower energy

report.

We use a pragmatic imputation approach based on the assumption that, given the stage of study completion, i.e., which fraction of the number of total recorded showers have been completed, showers by uploaders and non-uploaders have the same probability of having taken place after the first/second shower energy report. Formally, we assume that for each stage of study completion π,

$$Pr\left(IN_{it}^{s2} = 1 | \pi, non\text{-}uploader\right) = Pr\left(IN_{it}^{s2} = 1 | \pi, uploader\right).$$

To operationalize this approach, we estimate the distribution of uploaders' report timing over study completion non-parametrically, so $\widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi, uploader\right)$, and, instead of the indicator IN_{π}^{s2} for intervention stage 2, we define

$$\widehat{IN}_{s}^{s2} = \widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi_{it}^{s} = 1, uploader\right)$$

as probabilistic indicator for every shower of non-uploaders in study completion stage π. In other words, the regressor $\widehat{IN}_{\pi}^{s2}$ is the probability that a particular shower by a non-uploader took place after the first shower energy report. In all our regressions, we actually use the indicator

$$\widetilde{IN}_{it}^{s2} = \begin{cases} IN_{it}^{s2} & \text{if uploader} \\ \widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi, uploader\right) & \text{if non-uploader}. \end{cases} \tag{10}$$

Appendix E Supplementary Survey

We conducted a supplementary survey in a new sample of students in November and December 2019, about three years after the original experiment took place. The purpose of the survey was two-fold. First, we wanted to collect evidence that people tend to underestimate the environmental impact of showering without additional information. Second, we wanted to provide a manipulation check for our shower energy report intervention, testing whether the additional information on energy use and CO_2 emissions due to showering can plausibly induce stronger conservation efforts. The survey was conducted among residents of exactly the same student dorms in Bonn and Cologne in which the original study took place. Thus, the surveyee pool is comparable to the subject pool of the original experiment. In total, 329 students participated in the supplementary survey. Due to the high fluctuation rate of residents in student dorms, only 4 out of the 329 surveyees had also participated in the original experiment in 2016/17.

We first elicited students' prior beliefs about the amount of water used and CO_2 emitted per shower, as well as how confident they are about their response on a 10-point scale. As reference, we told surveyees that one hour of room lighting causes about 10

grams of CO_2 and that one hour of watching TV causes about 30 grams of CO_2. Furthermore, we asked students about their intention to take shorter showers on a 10-point Likert scale (we normalize this to mean 0 and standard deviation 1 for all analyses). After the first round of questions, we randomly presented one fact sheet (out of three) to each surveyee, mimicking the basic informational content of our original interventions. The "CON sheet" only contained information on average water temperature in the shower, the "RTF sheet" also included the average water use per shower, and the "SER sheet" further added information on energy use and CO_2 emissions. The exact wording was as follows. All fact sheets started with this text:

"Did you know that a few years ago, a study was conducted in this dorm, as well as other dorms in Cologne and Bonn? The study has shown that the average water temperature when taking a shower is about 37 degrees Celsius."

While the CON sheet ended here, the RTF sheet added the sentence *"... A typical shower uses around 40 liters of water."*. The SER sheet provided even more information by adding the following sentences: *"... A typical shower uses around 40 liters of water and 2.4 kWh of energy. This means that, on average, a persons emissions due to daily showering amount to almost 300 kg CO2 per year (800 grams per shower). It requires about 24 trees to absorb this amount of CO2."*. After surveyees had finished reading their respective fact sheet, we elicited posterior beliefs and attitudes by asking them the same questions again that they answered before receiving additional information. Surveyees were then paid 5 Euros for their participation in the survey, although 11 students refused to accept any remuneration.

Prior to receiving the fact sheets, surveyees estimated on average that they use 40.4 liters of water per shower (standard error of the mean = 6.36), causing emissions of 91.3 grams of CO_2 (s.e.m. = 15.03). While the estimate for water used per shower is roughly accurate on average, surveyees grossly underestimate the amount of CO_2 emitted by a factor of 8 to 9. However, subjects are also very uncertain about their estimates. On a scale from 1 (very uncertain) to 10 (very certain), the average surveyee places him-/herself at 4.24 for water use and 3.71 for CO_2 emissions.

Table A16 shows how surveyee change their beliefs and intentions after being provided with additional information through the fact sheets. Neither the RTF nor the SER survey induces statistically significant changes in surveyees' average estimates for water use per shower compared to the CON sheet, although surveyees in these groups become much more confident about their answer. In contrast, only the SER fact sheet has a strong impact on surveyees beliefs about CO_2 emissions. As surveyees severely underestimated the carbon intensity of showering in baseline, the SER fact sheet had an extreme debiasing effect compared to the CON and RTF fact sheets. This experimentally-induced belief update about environmental impacts is further associated with a sizeable increase in self-stated intentions to take shorter showers. Compared to surveyees receiving the RTF sheet, conservation intentions of surveyees receiving the SER sheet increased by 0.24 standard deviations ($p = 0.003$). In contrast, the RTF sheet did not increase intentions

Table A16: Supplementary survey — change in beliefs and intentions after fact sheet

	Water use per shower		CO_2 emissions		
	(1) Estimate	(2) Confidence	(3) Estimate	(4) Confidence	(5) Intention
RTF fact sheet	-12.274	2.148***	28.774	0.358*	0.060
	(10.146)	(0.279)	(21.587)	(0.208)	(0.065)
SER fact sheet	-22.909	2.561***	484.941***	2.023***	0.304***
	(16.813)	(0.258)	(37.599)	(0.264)	(0.076)
Constant	14.203	0.118	-15.274	0.335**	0.088**
	(9.663)	(0.161)	(19.023)	(0.138)	(0.042)
p-value for RTF = SER	0.451	0.175	0.000	0.000	0.003
Baseline mean	40.428	4.239	91.335	3.711	0.000
Observations	328	328	329	329	329
R^2	0.008	0.222	0.476	0.185	0.054

Robust standard errors in parentheses. The omitted category is the CON fact sheet group. Column (1) and (2) exclude one subjects who did not give a baseline estimate for water use. The intention measure used for column (5) is normalized to mean 0 and standard deviation 1.

significantly compared to the CON sheet ($p = 0.359$). Overall, these results suggests that people tend to severely underestimate the environmental impact of showering, and that information provision about energy and carbon intensity can induce subjects to increase their conservation efforts.

Appendix F More on Other Potential Mechanisms

F.1 Hawthorne or cueing effects

Given that we observe energy and water use in a relatively private and sensitive activity, showering, subjects' behavior may have been distorted by Hawthorne effects. We attempt to hold this constant by equipping every participant with a functioning smart shower meter, so to the degree that subjects in the control group respond to the sheer presence of a shower meter (with temperature feedback), we would in fact underestimate our conservation effects. To explain our empirical findings, Hawthorne effects would thus need to additionally interact with the intervention regimes. As the conservation effect in the RTF group (compared to the CON group) is quantitatively large and remains stable over the entire 3-months study duration, it seems unlikely that it is driven by differential Hawthorne effects. However, the shower energy reports may have made it more salient again to participants that they were part of a study, or alternatively, the reports may have simply served as a general cue or reminder to pay more attention to conservation efforts in the shower. Note that we sent out placebo emails instead of informative shower energy reports to the RTF and CON groups precisely to limit such types of

Table A17: Change in self-reported attitudes (baseline vs. post-intervention survey)

	shower comfort		*environmental attitude*	
	(1) ITT	(2) TOT	(3) ITT	(4) TOT
RTF group	0.042	0.047	-0.340***	-0.345***
	(0.117)	(0.119)	(0.117)	(0.119)
SER group	0.085	0.090	-0.277**	-0.253*
	(0.134)	(0.136)	(0.133)	(0.145)
DUAL group	-0.097	-0.011	-0.225*	-0.239*
	(0.138)	(0.150)	(0.129)	(0.144)
Constant	0.026	0.030	0.139	0.143
	(0.086)	(0.088)	(0.094)	(0.095)
F-test: p-value	0.641	0.896	0.034	0.039
Observations	300	255	304	257
R^2	0.007	0.003	0.027	0.031

Robust standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

confounders. Furthermore, we find that, if anything, the effect of shower energy reports (in the DUAL group) tends to become stronger over time instead of weaker, and the exercise in Figure 8 using different complier definitions for LATE estimation also suggests that it is the actual content of shower energy reports that matters. While we have no way to directly rule out Hawthorne or cueing effects, we are therefore confident that they do drive our empirical results.

F.2 Environmental attitude and consumption value of showering

Another alternative way in which two interventions could develop complementarities is through some sort of motivational spillover effect, in which the combined intervention convinced subjects to generally care more for the environment, or somehow made showering less pleasurable to them. Our interventions presented all information in a neutral and factual way, and we specifically refrained from including any normative element. Nevertheless, to check if this could confound our results, we again analyze subjects' survey responses before and after the study. The outcome variable of interest is the change in environmental attitude index or shower comfort index, respectively. All indices are normalized by subtracting the pre-intervention mean and dividing by the pre-intervention standard deviation.

The first two columns in Table A17 show difference-in-differences estimates for the effect of treatments on subjective shower comfort from baseline to endline survey. Both in the ITT (column 1) and in the TOT (column 2) regressions for subjective shower comfort, we find no significant differences across experimental condition, and all point estimates are virtually zero. Hence, at least based on self-reported measures, our interventions do not seem to have diminished the consumption benefits of showering, which is also

relevant for welfare considerations.

The other two columns in Table A17 show the difference-in-differences estimates for impacts on environmental attitude, with ITT estimates in column (3) and TOT estimates in column (4). Surprisingly, we find that subjects in the treated groups become *less* pro-environmental relative to the control group based on their survey responses. The magnitude of this decrease ranges from 22% to 35% of a (pre-study) standard deviation, which is not exactly quantitatively large, but also not negligible. We can only speculate about what is happening here. At face value, it may seem that feedback makes people less motivated to act pro-environmentally. Of course, we only have self-reported measures and cannot be certain of the underlying latent variable that they proxy for. But as we seem to proxy *self-perceived* inclination to act pro-environmentally rather than the actual extent of pro-environmental behavior, one possible interpretation could be that feedback provision curbs the capacity for distorted self-image formation, because people become aware of their intention-action gaps. We caution from overinterpreting the result here, as we did not have any ex ante hypothesis along these lines. Still, we can tentatively conclude that the conservation effects we observe are unlikely due to generally increased pro-environmental motivation.

Appendix G Theoretical framework: proofs and extensions

G.1 The Taylor approximation for policy complementarities in behavior

The second-order Taylor approximation for the change in behavior due to intervention X, Δe^X is given by:

$$\Delta e^X \approx \frac{\partial e}{\partial B} b_2 \delta_1 + \frac{1}{2} \frac{\partial^2 e}{\partial B^2} b_2^2 \left(\delta_1\right)^2 \tag{11}$$

and similarly for Δe^Y:

$$\Delta e^Y \approx \frac{\partial e}{\partial B} b_1 \delta_2 + \frac{1}{2} \frac{\partial^2 e}{\partial B^2} b_1^2 \left(\delta_2\right)^2 \tag{12}$$

The response to the bundle X, Y, Δe^{XY} is given by

$$\Delta e^{XY} \approx \frac{\partial e}{\partial B} \left[b_1 \delta_2 + b_2 \delta_1\right] + \frac{1}{2} \frac{\partial^2 e}{\partial B^2} b_2^2 \left(\delta_1\right)^2 + \frac{1}{2} \frac{\partial^2 e}{\partial B^2} b_1^2 \left(\delta_2\right)^2 + \left[\frac{\partial e}{\partial B} + \frac{\partial^2 e}{\partial B^2} b_1 b_2\right] \delta_1 \delta_2 \tag{13}$$

Thus, combining the last three expressions, we obtain

$$\Phi^{XY} \equiv \Delta e^{XY} - \Delta e^X - \Delta e^Y \approx \left[\frac{\partial e}{\partial B} + \frac{\partial^2 e}{\partial B^2} b_1 b_2\right] \delta_1 \delta_2 \tag{14}$$

As claimed in equation (5).

G.2 The Taylor approximation for policy complementarities in welfare

We first establish the derivatives of $W(b_1, b_2)$ as follows

$$\begin{aligned}\frac{\partial W}{\partial b_1} &= [V'(e) - c - \gamma]\frac{\partial e}{\partial B}b_2 \\ &= [(B-1)c - \gamma]\frac{\partial e}{\partial B}b_2\end{aligned} \tag{15}$$

$$\frac{\partial W}{\partial b_2} = [(B-1)c - \gamma]\frac{\partial e}{\partial B}b_1 \tag{16}$$

$$\frac{\partial^2 W}{\partial b_1^2} = b_2^2\frac{\partial e}{\partial B}c + [(B-1)c - \gamma]b_2^2\frac{\partial^2 e}{\partial B^2} \tag{17}$$

$$\frac{\partial^2 W}{\partial b_2^2} = b_1^2\frac{\partial e}{\partial B}c + [(B-1)c - \gamma]b_1^2\frac{\partial^2 e}{\partial B^2} \tag{18}$$

$$\frac{\partial^2 W}{\partial b_1 \partial b_2} = [(B-1)c - \gamma]\frac{\partial e}{\partial B} + b_1 b_2[(B-1)c - \gamma]\frac{\partial^2 e}{\partial B^2} + b_1 b_2 c\frac{\partial e}{\partial B} \tag{19}$$

We then calculate the second-order Taylor approximations of policy interventions X and Y as follows:

$$\begin{aligned}\Delta W^X &= \frac{\partial W}{\partial b_1}\delta_1 + \frac{1}{2}\frac{\partial^2 W}{\partial b_1^2}\delta_1^2 \\ &= [(B-1)c - \gamma]\frac{\partial e}{\partial B}b_2\delta_1 + \frac{1}{2}[b_2^2\frac{\partial e}{\partial B}c + [(B-1)c - \gamma]b_2^2\frac{\partial^2 e}{\partial B^2}]\delta_1^2 \\ &= [(B-1)c - \gamma]\Delta e^X + \frac{1}{2}b_2^2\frac{\partial e}{\partial B}c\delta_1^2\end{aligned} \tag{20}$$

where in the last equation, we substituted the definition of e^X from equation (11). Similarly, for ΔW^Y, we obtain

$$\Delta W^Y = [(B-1)c - \gamma]\Delta e^y + \frac{1}{2}b_1^2\frac{\partial e}{\partial B}c\delta_2^2 \tag{21}$$

The approximate welfare gain of the combined intervention (X, Y) can be expressed as:

$$\begin{aligned}\Delta W^{XY} &= \Delta W^{x} + \Delta W^{y} + \frac{\partial^2 W}{\partial b_1 \partial b_2}\delta_1\delta_2 \\ &= \Delta W^{X} + \Delta W^{Y} + [(B-1)c-\gamma](\frac{\partial e}{\partial B} + b_1 b_2 \frac{\partial^2 e}{\partial B^2})\delta_1\delta_2 + b_1 b_2 c \frac{\partial e}{\partial B}\delta_1\delta_2\end{aligned} \tag{22}$$

We finally use equation (13) to substitute $(\frac{\partial e}{\partial B} + b_1 b_2 \frac{\partial^2 e}{\partial B^2})\delta_1\delta_2 = \Delta e^{XY} - \Delta e^{X} - \Delta e^{Y}$ and obtain

$$\Delta W^{XY} - \Delta W^{X} - \Delta W^{Y} = [(B-1)c-\gamma](\Delta e^{XY} - \Delta e^{X} - \Delta e^{Y}) + b_1 b_2 c \frac{\partial e}{\partial B}\delta_1\delta_2 \tag{23}$$

as claimed in equation (7) in the main text.

G.3 Generalizing the complementarity condition to k dimensions of biases

In the case of $\prod_{k=1}^{K} b_k$ with $k > 2$, first we derive the k-dimensional geometric interpretation.

Suppose that b_1 is exogenously increased by δ_1. The resulting increase in B will be $\delta_1 \prod_{k=2}^{K} b_k$, as it is attenuated by b_2. Analogously, an exogenous increase of δ_j in the dimension of b_k results in an aggregate change of $\delta_j \prod_{k\neq j}^{K} b_k$. The effect of jointly increasing b_1 ,b_2 ... $b_k$by the same amounts, however, results in an overall change of

$$\Delta B = \prod_{k=1}^{K}\delta_k + \sum_{i=1}^{K} b_i \prod_{j\neq i}^{K}\delta_j + \sum_{i\neq j}^{i\leq K, j\leq K} b_i b_j \prod_{m\neq j, m\neq i}^{K}\delta_m + ... + \sum_{i=1}^{K}\delta_i \prod_{j\neq i}^{K} b_j. \tag{24}$$

In our context, k interventions X_1 , X_2 ... and X_k are complements if their combination reduces behavior by more than the sum of their individual effects, i.e $\Delta e^{\sum_{i=1}^{K} x_i} - \sum_{i=1}^{K}\Delta e^{X_i} < 0$.

$$\Delta e^{\sum_{i=1}^{K} x_i} - \sum_{i=1}^{K}\Delta e^{X_i} \approx \frac{\partial e}{\partial B}\sum_{i\neq j}^{i\leq K, j\leq K}\frac{\prod_{m=1}^{K} b_m}{b_i b_j}\delta_i\delta_j + \frac{\partial^2 e}{\partial B^2}\sum_{i\neq j}^{i\leq K, j\leq K}\frac{\prod_{m=1}^{K} b_m^2}{b_i b_j}\delta_i\delta_j$$

However, if our "Anna Karenina" condition is satisfied, i.e., if any terms among b_k are sufficiently close to zero, the first term always dominates, and complementarities in bias reduction translate into complementarities in behavior.

As for the welfare,

$$\Delta W^{\sum_{i=1}^{K} x_i} - \sum_{i=1}^{K}\Delta W^{X_i} \approx [(B-1)c-\gamma]\left(\Delta e^{\sum_{i=1}^{K} x_i} - \sum_{i=1}^{K}\Delta e^{X_i}\right) + \sum_{i\neq j}^{i\leq K, j\leq K}\frac{\prod_{m=1}^{K} b_m^2}{b_i b_j} c\frac{\partial e}{\partial B}\delta_i\delta_j$$

If, in addition, any terms among $\Delta e^{X_i} \approx 0$, this implies that one of the b's is equal to zero, and thus complementarities also exist in terms of welfare.

G.4 Generalizing the complementarity condition to other aggregations of bias

If $\delta_1^{X+Y} = \delta_1^X + \delta_1^Y$ and $\delta_2^{X+Y} = \delta_2^X + \delta_2^Y$, The change in behavior is

$$\Delta e^{X+Y} - \Delta e^X - \Delta e^Y = \frac{\partial^2 e}{\partial B^2}\left(\delta_1^X\delta_1^Y b_2^2 + \delta_2^X\delta_2^Y b_1^2 + \delta_1^X\delta_2^Y b_1 b_2 + \delta_2^X\delta_1^Y b_1 b_2\right) + \frac{\partial e}{\partial B}\left(\delta_1^X\delta_2^Y + \delta_1^Y\delta_2^X\right). \tag{25}$$

If $\delta_i^{X+Y} = \max(\delta_i^X, \delta_i^Y)$, for example $\delta_1^X = \max(\delta_1^X, \delta_1^Y)$, $\delta_2^Y = \max(\delta_2^X, \delta_2^Y)$ The change in bahavior is

$$\begin{aligned}\Delta e^{X+Y} - \Delta e^X - \Delta e^Y = \frac{\partial^2 e}{\partial B^2}&\left(\delta_1^X\delta_2^Y b_1 b_2 - \delta_1^X\delta_2^X b_1 b_2 - \delta_1^Y\delta_2^Y b_1 b_2 - \frac{1}{2}\delta_1^{Y^2} b_2{}^2 - \frac{1}{2}\delta_2^{X^2} b_1{}^2\right) \\ &+ \frac{\partial e}{\partial B}\left(\delta_1^X\delta_2^Y - \delta_1^X\delta_2^X - \delta_1^Y\delta_2^Y - \delta_1^Y b_2 - \delta_2^X b_1\right).\end{aligned} \tag{26}$$

To illustrate the complementarity in an example that might be closer to reality, consider the case of two sources of bias and two interventions, X and Y. Suppose that intervention X is primarily targeted at the perception of the environmental impact b_1, while potentially also having a positive side-effect on b_2, which could describe an information intervention which may also lead to endogenously higher attention levels (Hanna et al., 2014; Gabaix, 2017). Analogously, intervention Y is primarily targeted at the attention parameter b_2, with positive side-effects on b_1. This could describe a salience intervention that incidentally also offers some degree of information or induces information search efforts. Hence, the relevant parameters are such that $\delta_1^X \geq \delta_1^Y$ and $\delta_2^Y \geq \delta_2^X$. The reduction in bias of each intervention in isolation are $\Delta B^X = \delta_1^X b_2 + \delta_2^X b_1 + \delta_1^X\delta_2^X$ and $\Delta B^Y = \delta_1^Y b_2 + \delta_2^Y b_1 + \delta_1^Y\delta_2^Y$, respectively, which is also illustrated in Figure A2a and b.

Aggregating policy interventions. — When two partially overlapping interventions are introduced jointly, we need to specify how they aggregate into the overall bias B. As a benchmark, we assume that the mitigation effects δ_i^X, δ_i^Y are additive (and that the resulting b_i does not exceed 1). Figure A2c illustrates this example, in which $\delta_1^{X+Y} = \delta_1^X + \delta_1^Y$ and $\delta_2^{X+Y} = \delta_2^X + \delta_2^Y$. The additional bias reduction is

$$\Delta B^{X+Y} - \Delta B^X - \Delta B^Y = \delta_1^X\delta_2^Y + \delta_2^X\delta_1^Y. \tag{27}$$

Notice, that — holding constant δ_1^{X+Y} and δ_2^{X+Y} — the potential for complementarity is largest for two completely specialized interventions.

Next, we look at a case where, in each dimension, only the dominant intervention matters, i.e., $\delta_i^{X+Y} = \max(\delta_i^X, \delta_i^Y)$ c. This is illustrated in Figure A2d. This case is less favorable toward complementarities, as each intervention now only has an impact on

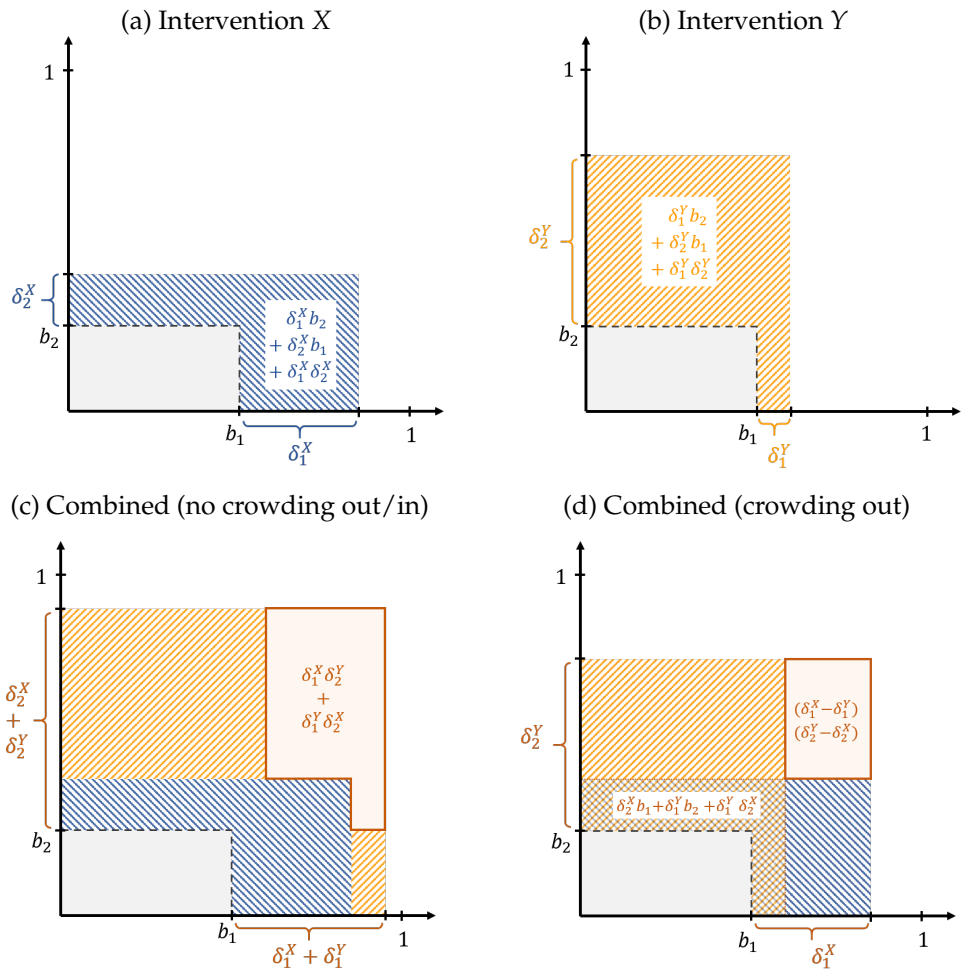


Figure A2: Depiction of example interventions

Notes. Figures (a) and (b) illustrate the bias mitigation effect of interventions X and Y in isolation, respectively. Figure (c) illustrates their combined effect when their individual effects in each dimension are additive, i.e., there is neither crowding out nor crowding in. Figure (d) illustrates their combined effect when there is perfect crowding out of the less effective intervention in each dimension.

one bias dimension, and the condition becomes

$$\Delta B^{X+Y} - \Delta B^X - \Delta B^Y = (\delta_1^X - \delta_1^Y)(\delta_2^Y - \delta_2^X) - (\delta_2^X b_1 + \delta_1^Y b_2 + \delta_2^X \delta_1^Y) \tag{28}$$

This term is positive if the top right rectangle in Figure A2d, which represents the policy lever complementarity, is larger than the cross-shaded intersection of X and Y, which represents loss in impact from X and Y in isolation. Complementarity is more likely the more specialized each intervention is, as the interaction is increasing in b_1^X and b_2^Y and decreasing in b_1^Y and b_2^X.