Essays in Applied Microeconomics

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch

die Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Ximeng Fang

aus Dortmund

Bonn, 2022

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Lorenz Goette
Zweitreferent:	Prof. Dr. Thomas Dohmen
Tag der mündlichen Prüfung:	29. April 2022

Acknowledgments

First and foremost, I would like to thank my supervisors Lorenz Goette and Thomas Dohmen, who have been guides and inspirations to me since my first steps into the world of research, and from whose generous advice and vast experience I continue to benefit tremendously. Although I still find it rather funny, I do understand now why the German name is literally "doctoral father". I am also incredibly grateful for the opportunity to work with my wonderful coauthors and for help I received from other researchers at various occasions. Special thanks go out to Verena Tiefenbeck, who has been a cheerful and dedicated collaborator on many projects, and to Matthias Sutter, who never hesitated to provide recommendation letters for me.

My gratitude also goes to Andrea Reykers, Simone Jost, Ilona Krupp, Silke Kinzig, and the entire administrative staff of the IAME, the BGSE, the CRC TR 224, and the ECONtribute, who have supported me and my research continuously over the past years and somehow always managed to find answers and solutions to whatever bizarre questions and requests I had. Thanks also to all the student helpers who have made my research possible through their tireless work in the field.

Furthermore, I would like thank my friends and fellow PhD students, who have made this entire journey so much more enjoyable and fun. Here's to you(!): Jing Jing (the best partner-in-crime imaginable); Chui Yee (Asianness in a small (but oh my!) nutshell); Tobias (the flagbearer of SCT and the Mensa); Timo (although he supports the wrong football club); Matthias (he supports the right one); the briq-brigade Felix, Luca (either entertaining or asleep, usually both), and Peter; Johannes S., Sven, Lasse, and Johannes W. (with all of whom I have not only shared offices, but also plenty of laughs, and even the occasional talk about research); Sonja (for fun coffee/lunch/Gin breaks); Alina (for Rohan!); the fellow student reps Lukas, Ricardo, and Radost ("we go way back"); Janos; Klara; Christian A.; Christian Z.; Anna (who always made short work of my leftover cake experiments); Puja; Hanna and Sascha (don't DIE); Elisa; Freddy, Lars, Marie, Nadja, and Theresa (who have all played a part in nudging me into this whole thing); Florian, Malte, Nina, Steffi, and Tim (for being the best gang ever and coming up with "Dr. X"); and many more.

Finally, I thank my parents for their unconditional love, support, and confidence in me, without which I would never have gotten anywhere far in life.

Contents

List of Figures

List of Tables

Introduction

"The noblest pleasure is the joy of understanding." — Leonardo da Vinci

We as humankind stand at a crossroads, with an unprecedented number of opportunities ahead of us due to technological progress and wealth of information, yet also facing existential challenges such as climate change, a deadly pandemic, social and political polarization. One common thread is that it is us, our own actions and inactions, that will determine the future outcomes for ourselves, our society, and our planet. Thus, understanding human behavior — with all its facets and intricacies —, and how to change it for the benefit of ourselves and others, has never been as pressing as it is today. While our lives have grown ever more complex and interconnected, it is in light of this overwhelming complexity that our own cognitive biases and limitations, and the limits to our understanding of the people and the world around us, become all the more evident. This thesis — consisting of four independent chapters — aims to further our understanding both of the nature of human behavior as well as of how we may (or may not) leverage these insights to enact simple policy interventions that help individuals and groups to act in a way that is beneficial for society and in line with their own values and intentions.

One of the greatest threats in the 21st century is posed by climate change and resource scarcity. Amidst growing public concern about these issues, many individuals are willing to make personal sacrifices to protect the environment; yet, we often observe a gap between intentions and actions. This gap may be partly driven by behavioral frictions and biases such as, e.g., imperfect information, limited attention, self-control problems. Understanding these limitations and how to overcome them could contribute to demand-side approaches to resource conservation that can complement technological solutions and conventional policy instruments like carbon taxation (Dietz, Gardner, Gilligan, et al., 2009; Allcott and Mullainathan, 2010; Creutzig, Roy, Lamb, et al., 2018).

Modern digital technologies enable the use of personalized interventions in a large variety of new applications by allowing for precise quantitative measurement of consumer behavior. **Chapter 1: "Goal-Setting and Behavioral Change"** of this dissertation examines the effects of goal-setting interventions through modern smart metering tools in the context of household water conservation in Singapore, one of

the most water-stressed countries in the world. While psychologists have long argued that (challenging yet attainable) goals can motivate effort even when they there are no material consequences tied to goal success or failure, economists have only recently begun considering how to incorporate such nonbinding goals into their decision-making and policy intervention frameworks. We provide causal evidence on the effects of externally set goals and real-time feedback on water conservation in the shower from a randomized field experiment with 525 households and over 2,000 individuals, using smart meters to collect fine-grained behavioral measures continuously over a duration of 4 to 6 months. Our results provide strong evidence that goal-setting can lead to significant average conservation effect on top of real-time feedback, thus constituting a simple and scalable intervention tool. Based on fine-grained analyses of the around 300,000 shower observations in our sample, we further find that the impact of goals is mostly "local": effects are particularly strong when a goal is in sight, but quickly dissipate when it becomes out of reach. This suggests a discontinuous jump in the utility function at the goal — akin to a warm glow effect. Interestingly, goals seem to become less meaningful over time to individuals, as bunching and goal attainment rates gradually decline, although average conservation effects remain stable, which is consistent with nonbinding goals taking on the role of initial norms for acceptable consumption levels.

Many other empirical studies have demonstrated how non-monetary behavioral interventions can induce household energy and water conservation through the use of a variety of tools to overcome behavioral frictions and biases. However, what is less well understood is how different types of interventions interact with each other, and in particular which behavioral mechanisms may systematically induce complementarity or substitutability between interventions. That is the research question in **Chapter 2: "Complementarities in Behavioral Interventions"**. In this study, we argue that when multiple behavioral barriers operate simultaneously, interventions that each target different barriers may be complements, i.e., each intervention becomes more effective when combined with the other. For example, if consumers tend to underestimate the environmental impact of their actions and additionally suffer from behavioral biases like inattention, then it may be necessary to address both issues at the same time. We implement this idea empirically in a randomized field experiment (n = 351) on energy and water conservation in a resource-intensive everyday activity, again showering, combining two different behavioral intervention in a 2×2 design. The first intervention, shower energy reports, primarily aimed at improving knowledge about environmental impacts of warm water consumption; the second intervention, real-time feedback, primarily aimed at increasing salience of resource use in a simple and timely manner. Our empirical results show that, in isolation, the first intervention had no statistically significant impact on energy consumption, despite inducing strong knowledge gains, whereas the latter consistently reduced energy consumption by 17-18% on average throughout the entire three-month study period. Crucially, implementing both interventions

together boosted the conservation effect of real-time feedback in isolation by over 50%, pointing towards a striking complementarity. Thus, our evidence suggests that in situations where multiple behavioral frictions and biases play a role, appropriate policy bundling may be necessary for behavioral policy to unfold its true potential.

The Covid-19 pandemic poses another potentially disastrous challenge to our society, having already cost millions of lives and caused vast social and economic mayhem in a time span of just two years. Digital contact tracing (DCT) apps can provide critical relief for public health authorities in breaking infection chains and slowing the spread of the virus, if adopted by a sufficiently large share of the population. **Chapter 3: "Motivating the Adoption of Digital Contact Tracing Apps"** presents results from a large-scale social media experiment (n $\approx$ 1 million impressions) aimed at encouraging more people to install and use DCT apps in the second wave of the Covid-19 pandemic in Germany. Our intervention provides feedback and social comparison on regional Covid-19 incidence rates through a video advertisement that also incorporates a regular promotion clip for the Corona Warn App (CWA, the official German DCT app). Providing locally targeted feedback increases video plays and click-through rates (i.e. following a link to the CWA homepage) by 30% compared to a control condition that only shows the regular promotion video. Highlighting that the incidence rate in the county of residence is higher compared to other counties increases ad views and click-through rates by an additional 15%. However, we observe an overall negative relationship between the effectiveness of the treatment and local incidence rates, which is most likely driven by county characteristics that determine both incidence rates (i.e. through lower compliance with social distancing) and interest in our interventions. We additionally replicate the experiment in a representative online survey (n $\approx$ 6,000) and observe small but statistically significant increases in positive attitudes towards the DCT app in response to locally targeted feedback on incidence rates. Unfortunately, the survey indicates a low baseline willingness to adopt the CWA, thus casting doubt on the persuasive power of simple behavioral interventions beyond a small subset of individuals at the margin.

The previous three chapters have focused mostly on individual decision-making without explicitly taking into account the social environment in which it is embedded. However, humans are social animals. What we do, and how we do it, is fundamentally determined by our drive to conform to social norms and our desire to maintain a positive image in front of peers as well as strangers. This implies that making individuals' actions publicly visible may spur them to act in a way that presents them in the best possible light to others. In **Chapter 4: "The Effect of Transparency on Performance Evaluation in Committees"**, we study the effects of public visibility on decision-making by committees of experts in a professional context. High-stakes decisions or recommendations are often delegated to committees rather than single experts in order to collect more information and to diversify opinions. However, the quality of decision-making in a committee can depend on institutional features

such as whether inputs of each member are made public or kept secret. We study the effects of a transparency reform in competitive figure skating on performance scores awarded by the panel of judges, and we attempt to understand the empirical results through the lens of a beauty-contest model in which transparency influences evaluation decisions through increased conformity concerns. Using a difference-in-differences design and seven seasons of data from almost 17,000 performances, we show that the dispersion of (artistic) scores within a judge panel decreased significantly after the reform, indicating a larger degree of consensus and potentially higher judge effort. This effect is stronger for high-profile competitions, which could be due to higher levels of anticipated public scrutiny. However, we also find that the reform did not result in a lower aggregate degree of nationalistic favoritism, contrary to its original intention.

One common thread throughout this dissertation is the desire to understand how the observed actions of individuals and groups respond to seemingly small changes in their everyday decision environments — simple information, nonbinding goals, salience of certain attributes, public visibility. They have illustrated both the potential and the limits of such interventions, and they have searched for the answer to "why?" in the many facets of human motivation. My hope is that the continued striving for a better understanding of our social and economic behavior will not only bring pleasure and joy, but also generate valuable insights that may help us in building a better future for ourselves and society at large.

References

Allcott, Hunt, and Sendhil Mullainathan. 2010. "Behavior and Energy Policy." *Science* 327 (5970): 1204–1205. [1]

Creutzig, Felix, Joyashree Roy, William F. Lamb, Inês M. L. Azevedo, Wändi Bruine de Bruin, Holger Dalkmann, Oreane Y. Edelenbosch, Frank W. Geels, Arnulf Grubler, Cameron Hepburn, Edgar G. Hertwich, Radhika Khosla, Linus Mattauch, Jan C. Minx, Anjali Ramakrishnan, Narasimha D. Rao, Julia K. Steinberger, Massimo Tavoni, Diana Ürge-Vorsatz, and Elke U. Weber. 2018. "Towards demand-side solutions for mitigating climate change." *Nature Climate Change* 8 (4): 260–263. [1]

Dietz, Thomas, Gerald T. Gardner, Jonathan Gilligan, Paul C. Stern, and Michael P. Vandenbergh. 2009. "Household Actions Can Provide a Behavioral Wedge to Rapidly Reduce US Carbon Emissions." *Proceedings of the National Academy of Sciences of the United States of America* 106 (44): 18452–18456. [1]

Chapter 1

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

Joint with Sumit Agarwal, Lorenz Goette, Tien Foo Sing, Verena Tiefenbeck, Samuel Schoeb, Thorsten Staake, and Davin Wang

1.1 Introduction

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

Goal-setting is a simple and popular motivational tool. A large body of literature in psychology has demonstrated the motivating power of goals even when they are nonbinding, i.e. there are no explicit material rewards tied to achieving or failing the goal (Locke and Latham, 1990; Locke and Latham, 2002). Similarly, the notion of "Management by Objectives" (Drucker, 1954) has been highly influential in both the theory and practice of organizational management. While economists have long studied the use (monetary) bonus incentives in organizations, they have only

recently begun exploring the role of payoff-irrelevant goals, mostly in the context of self-set goals that agents can use as soft commitment tools against self-control problems (e.g. Koch and Nafziger, 2011; Harding and Hsiaw, 2014; Allen, Dechow, Pope, et al., 2017; Clark, Gill, Prowse, et al., 2020).[1]

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

In this paper, we provide causal evidence on the effects of goal-setting and real-time feedback from a randomized field experiment with over 2,000 individuals from 525 households in the context of everyday water conservation, using smart meters to continuously collect fine-grained behavioral measures over a duration of 4 to 6 months. We conducted the experiment in Singapore, a severely water-stressed country, where government agencies have made it a high priority to reduce daily domestic water consumption per capita to 130 liters by 2030 (down from 141 liters in 2018), for example by promoting a wide range of water savings campaigns, often stressing that "every drop counts".[2] In our study, we target a particularly water-intensive activity, namely showering, which constitutes almost 30% of total water usage in Singaporean households (PUB, 2018a). All households were equipped with Amphiro smart shower meters that were directly installed in the shower and that automatically recorded detailed information on water usage patterns every time the shower is used. Overall, we collected data from about 320,000 shower observations over the entire course of the study.

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

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

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

tervention. Thus, we have experimental treatment variation both across and within subjects.

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

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

Overall, the empirical results show that our interventions have a strong motivating effect on water conservation behavior. Consistent with earlier studies, we find that real-time feedback alone already leads to significant reductions in average water usage by 1.87 liters per shower relative to the Control group, which corresponds to an effect size of about 9 percent. Importantly, externally set goals can increase conservation efforts dramatically, the reductions being twice as high (3.92 liters per shower) in the 15L condition – which turned out to be the most effective

of all Goal conditions based on point estimates. However, we also find that the easiest (35L) did not lead to any additional reductions in water usage compared to real-time feedback alone, with the estimated conservation effect of 1.11 liters even being somewhat smaller. In addition, the relation of goal difficulty and effort appears to be non-monotonic: while the 20L and 25L goal led to a reduction of around 3.0 liters per shower on average, the point estimate for the most ambitious goal (10L) is 2.97 liters and thus smaller than the one for the moderately ambitious 15L goal. This non-monotonic pattern of the point estimates bears close resemblance to the conventional notion that the best goals are challenging yet attainable (Locke and Latham, 1990; Heath, Larrick, and Wu, 1999).

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

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

In addition to investigating differences in (conditional) average outcomes across experimental conditions, we make use our sample of the around 300,000 shower observations to further examine more fine-grained behavioral responses as a function of distance to the experimentally assigned goal. Based on the RTF and Control conditions, we can further construct experimental placebos to compare outcomes with and without an exogenous goal, which offers methodological advantages to studies that rely on smoothness assumptions about the counterfactual distribution Allen et al. (e.g. 2017). We observe strong bunching at goals, with the share of show-

ers in the 0.5 liter bin below a conservation target being about 16% higher than the corresponding share in the RTF group. Using non-parametric survival analysis methods, we find that hazard rates of showers are mostly affected locally around a goal. As the amount of water used in the shower approaches the goal, the stopping rate gradually increases, peaking in the last moments before they would fail the goal. Intriguingly, we observe a sharp upward spike in the stopping probability by 44% at the very last deciliter below it. However, after the water volume has surpassed the conservation target, the hazard rate quickly drops to the level of the RTF group. This pattern of stopping probabilities strongly suggests that individuals experience a discontinuous jump in utility depending on whether they achieve the goal or not, which may be interpreted as psychological bonus reward or warm glow. In contrast, it is inconsistent with frequently used models of goals as reference points that induce loss aversion (e.g. Heath, Larrick, and Wu, 1999; Koch and Nafziger, 2011; Gómez-Miñambres, 2012), as this would predict persistently higher hazard rates once the goal is surpassed.[3]

Finally, we investigate whether the motivational power of goals is short-lived or remains stable over time. For instance, it may be the case that individuals simply become numb towards attainment or failure of nonbinding goals after some time, e.g. due to disengagement after repeated failure (Höpfner and Keith, 2021). In contrast, we find that the average conservation effects of all treatments are remarkably stable over time, with no evidence of waning (or strengthening) over a period of 4 to 6 months. Seemingly at odds with this finding, we also observe a significant decrease in bunching and goal attainment rates over the course of the study, which indicates that individuals develop a more nonchalant attitude towards the specific goal assigned at the beginning of the study. Thus, externally set goals seem to serve as norms or default for an acceptable level of water usage per shower, with repeated experience leading individuals to form habits or adjust their expectations.

Our paper contributes to the growing literature on demand-side approaches to promote pro-environmental behavior. Behavioral interventions aimed at overcoming such barriers have been used to facilitate behavioral change in a variety of contexts such as retirement savings or public health (Thaler and Sunstein, 2008; Madrian, 2014), and are also regularly advocated as promising policy tool for fostering more environmentally sustainable household consumption behavior (e.g. Dietz, Gardner, Gilligan, et al., 2009; Allcott and Mullainathan, 2010; Reddy, Montambault, Masuda, et al., 2017; Creutzig, Roy, Lamb, et al., 2018).[4] For example, influential early studies have demonstrated the impact of social-norm based interventions on

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

4. Pro-environmental interventions have drawn from a broad set of instruments such as information provision, social norms, goal-setting, etc. While the general findings are that non-monetary interventions can be an effective tool in reducing energy and water usage, the quantitative effect size

household energy and water conservation Allcott (e.g. 2011), Ayres, Raseman, and Shih (2013), and Ferraro and Price (2013a). While these interventions typically provide feedback on aggregate household consumption, recent studies have argued that interventions that enable better behavioral control and learning, e.g. through activity-specific disaggregation (Gerster, Andor, and Goette, 2020) and higher frequency (Allcott and Rogers, 2014; Tiefenbeck, Goette, et al., 2018), may increase the effectiveness. For example, in a closely related studies, Tiefenbeck, Goette, et al. (2018) provide activity-specific real-time feedback in the shower through the same type of smart meter that we use in this study and document a 22% conservation effect, or, in absolute terms, a reduction of 0.6 kWh energy and 9 liters of water per shower. These results also replicate in a sample without monetary incentives and without self-selection into the study (Tiefenbeck, Woerner, Schoeb, et al., 2019). A natural question that we address is whether technology-based feedback interventions, enabled by advances in digitization and smart metering, can be enhanced by including further motivational tools like goal-setting.

Decades of studies in psychology have demonstrated the potential of nonbinding goals (or "mere" goals) for improving task performance in a large variety of contexts (Mento, Steel, and Karren, Ronald, J., 1987; Locke and Latham, 1990, 2002; Locke and Latham, 2019b). While economists have recently begun exploring the use of goal-setting for example to motivate healthy food choice (Samek, 2019), student performance (Dobronyi, Oreopoulos, and Petronijevic, 2019; Clark et al., 2020), worker effort (Corgnet, Gómez-Miñambres, and Hernán-González, 2015; Brookins, Goerg, and Kube, 2017; Fan and Gómez-Miñambres, 2020), or energy conservation (Abrahamse, Steg, Vlek, et al., 2007; Harding and Hsiaw, 2014), there is no clear guidance yet how to incorporate nonbinding goals into economic decision-making frameworks. We contribute to the literature on goal-setting by providing field evidence from a randomized experiment in a diverse sample with continuous measurement of behavior over an extended period of 4 to 6 months. While our results are consistent with many previous findings from the psychology literature — in particular that goals can motivate effort provision if they are challenging and attainable —, we further contribute to the understanding of goal-directed behavior by collecting a large data set of about 300,000 measured observations and examining fine-grained behavioral patterns in response to different goals. In line with Allen et al. (2017), who document discontinuities in the distribution of marathon finish times at round numbers (e.g. 3h, 3:30h, ...), we observe bunching of water volumes at the goal. Compared to Allen et al., our study offers methodological advances by experimentally assigning different goals to subjects and their comparing outcomes to subjects who did not receive any explicit goal. We further contribute to the literature by pro-

may be relatively small (around 2%) on average in methodologically more rigorous studies. For reviews, see e.g. Abrahamse, Steg, Vlek, et al. (2005), Fischer (2008), Delmas, Fischlein, and Asensio (2013), Karlin, Zinger, and Ford (2015), Andor and Fels (2018), Carlsson, Gravert, Kurz, et al. (2021).

viding evidence on nuanced dynamic effects of repeated everyday exposure to a goal for several months.

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

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

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

6. Evidence on the "joy of winning" (Dohmen, Falk, Fliessbach, et al., 2011) as well as models of aspiration levels (Diecidue and van de Ven, 2006) also consider discrete psychological rewards attached to a binary representation of success.

1.2 The empirical setup

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

1.2.1 Sample recruitment and study procedures

We recruited household from public housing blocks (HDBs) in 27 geographical nodes with varying population density and composition that are dispersed over the entire island and selected to create a broadly representative cross-section of Singaporean households. Appendix Figure 1.A.1 shows the geographical distribution of the participating HDB sites across the island. Note that 80% of the Singaporean population lives in HDB apartments that are built and sold by the Housing Development Board.[7]

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

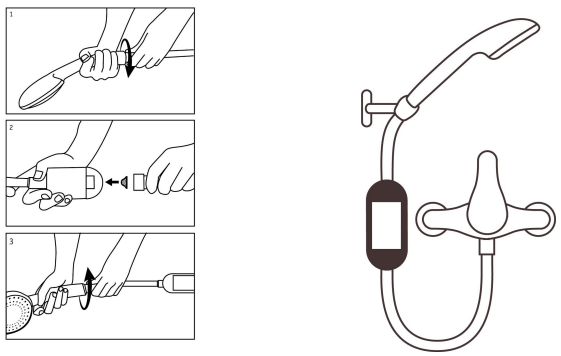
We distributed smart meters to all participating households to measure their water usage in the shower and to deliver the real-time feedback and goal-setting interventions. Deployment of the devices was carried out in June and July 2015 and the regular study duration was four months, with a subset of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

1.2.2 The feedback and measurement technology

All participating households were equipped with one Amphiro a1 smart shower meter for each bathroom in their apartment. The smart meter could be easily installed

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

8. While direct retrievals were preferred, because we could check if devices were still installed and get a feeling of participants' attitudes, not all of them could be reached easily and we arranged for device retrieval via postal service for 25 households.



(a) Device installation **(b)** Installed device

Figure 1.1. Position of the Amphiro a1 smart meter in the shower.

between the shower head and the shower (see Figure 1.1), after which it measured and recorded water usage variables of every shower taken. It is small, lightweight, and powered by an integrated hydro turbine that does not noticeably affect water flow. Furthermore, it features a smartphone-sized liquid crystal display, which we programmed to show various types of information tailored specifically for the purposes of this study.

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

We define "showers" as water extractions of at least 4.5 liters volume with an average flow rate of at least 2 liters per minute, whereas we classify observations as non-shower water extractions otherwise. The smart meters only stored detailed usage data for observations that qualify as showers according to these criteria. The reason for this restriction is that the storage capacity of each smart meter allowed for a maximum of 672 data points, which was in fact reached by 16% of the study

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

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

Figure 1.2. Amphiro a1 display in different configurations

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

devices. Therefore, we wanted to avoid wasting storage space for minor water extractions, e.g. for cleaning purposes. However, we did program the devices to count the total number of all water extractions as well as the cumulative amount of water used, also including extractions that did not qualify as showers. Similarly, the device stored the number of showers from the 673rd showers onward, as well as the average water volume used for these showers.

1.2.3 Experimental conditions

The smart meter could display tailored pieces of information to participants in real-time, i.e., while they were using their shower. For the purposes of our study, we programmed the smart meters to be in one of three possible modes, depending on the study progress and the assigned treatment. In mode 1, the device only displayed information on the current water temperature (in degrees Celsius); in mode 2, it provided real-time feedback on the absolute amount of water used for the current shower; in mode 3, it additionally provided relative feedback on water usage vis à vis a fixed conservation goal. The different display modes are illustrated in Figure 1.2.

Households were randomly assigned to one of seven experimental conditions and received smart meters in the respective configurations. In the control condition, subjects only ever saw information on water temperature while showering. In the real-time feedback only (RTF) condition, subjects' had devices that also displayed real-time feedback on current water usage, but did not include a conservation goal. This treatment closely resembled the ones in Tiefenbeck, Goette, et al. (2018). Furthermore, there were five different Goal conditions, in which subjects received devices that incorporated an exogenously assigned volume goal in addition to real-time feedback on absolute water usage. The goals were set at 10L, 15L, 20L, 25L, and 35L, respectively, and subjects were encouraged to keep their water consumption

below this amount.[10] No explanations for the choice of the goal were provided, and the specific goal level was only revealed with the 21st shower, when the intervention period began. From then on, it was displayed during the first ten seconds of the shower on the LCD. During showering, the display showed an injunctive message that rated the current water consumption level as “very good" if it was below 7 liters, "okay" if it was above 7 liters but below the respective conservation goal, and "too much" if it exceeded the goal.

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

1.3 Data and descriptive statistics

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

1.3.1 Water usage data

The smart meters recorded information on, among others, the water volume, water temperature, and time duration of all showers taken during the study. All but 2 of the 884 study devices we had deployed could be retrieved from the households, but for 41 devices we were not able to read out any valid data despite multiple attempts, potentially due to defective storage. Furthermore, 14 devices had no data stored at all, probably because they were never used by the households. We also have to exclude 3 households to which we had accidentally sent wrongly configured devices. Overall, we obtained valid water usage data for about 320, 000 recorded shower observations from 822 devices and 511 households, representing over 2, 000 individu-

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

Table 1.1. Number of observations by experimental condition

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

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

als.[11] Table 1.1 provides an overview of the number of observations by experimental condition.

For most of our analyses, we only include devices that have recorded more than 20 shower observations, as devices with 20 observations or fewer stayed in baseline mode for the entire study and do not help us in empirically identify the effect of our interventions. Excluded devices have most likely been installed in bathrooms that are very infrequently used. Table 1.1 shows that this restricted analysis sample contains data from 774 devices, with the number of shower observations remaining close to 320, 000. Out of these observations, 28, 493 showers were recorded after the device had reached its storage limit of 672 data points. For these showers, we do not observe individual measures of water usage, but instead of this we observe the average water volume of all post-limit showers registered by a device. If not stated explicitly otherwise, we will also make use of these imputed observations for analyzing impacts on average water usage per shower.

1.3.2 Survey data

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

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

on general environmental attitude, shower comfort, and perceived water consumption ("How many liters of water do you think you use per shower?"). The response rate for the baseline survey was 99%.

In addition, households were asked to complete a short online follow-up survey two months after device installation, and households with study duration of six months completed an identical online survey again two months later. Finally, we conducted an in-person endline survey when retrieving the devices, whenever possible with the same individual who completed the baseline survey.[12] The follow-up and endline surveys contained questions on experiences with the shower meter, such as whether participants believed that it was helpful, stressful, effective in changing showering behavior, and whether the goal was too difficult. In addition, they included the same set of questions about attitudes and beliefs towards water usage and water conservation as in the baseline survey. More than 95% of the households completed the follow-up and endline surveys.

1.3.3 Household characteristics

Descriptive statistics on household and participant characteristics are presented in Table 1.2 and compared to the general Singaporean population in HDB dwellings.[13] As we recruited our sample mostly from larger HDBs, the average household in our study consists of 4.2 members, while the average household size in HDBs in Singapore was 3.34 in 2015. The apartment of a modal household contained four to five bedrooms and two bathrooms. 79% of the participating households are ethnic Chinese and 12% are ethnic Indians, whereas ethnic Malay households form 5% of our sample. The composition is roughly representative of the population in Singapore, albeit with an underrepresentation of Malays relative to Chinese, Indians, and Others. The average age of individuals from participating households in our sample was 36.5 (median 35) — compared to the HDB population average of 37.9 in Singapore. About 17% of the participants were below age 15, and 10% were 65 years old or higher. Thus, our sample spans all age groups, sometimes comprising three generations within the same household, which is not uncommon in Singapore. The female share among our subjects was 53%.

1.3.4 Number of showers and water extractions

On average, we observe about 414 showers per bathroom over the entire 4 (to 6) months period of the trial, which corresponds to a frequency of approximately 1.3

12. 25 households sent their devices back via postal service, as we could not find a suitable retrieval appointment. In these cases, the final survey was either conducted over the phone or they filled out a paper-based survey instead.

13. Source: Department of Statistics Singapore (singstat.gov.sg).

Table 1.2. Sample characteristics

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

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

recorded showers per person every day.[14] One concern about our intervention may be that individuals compensate shorter showers with more showers or, vice versa, that they avoid showering and thereby compromise basic hygiene needs. Furthermore, we may overestimate effects of our intervention on overall water consumption if individuals partially relocate water usage from the private shower to other facilities (e.g. wash basin, gym showers). To alleviate these concerns, we compare the total recorded number of showers per bathroom across experimental conditions in Figure 1.3. There is no evidence for differences in the number of showers ($p = 0.9682$). We confirm this in further robustness checks in Appendix Table 1.B.2.

14. We calculate this using information on the number of all household members (including potentially non-permanent members) reported in the baseline survey and information on the dates of installment and retrieval for each smart meter. The net frequency adjusted for absences may be even higher.

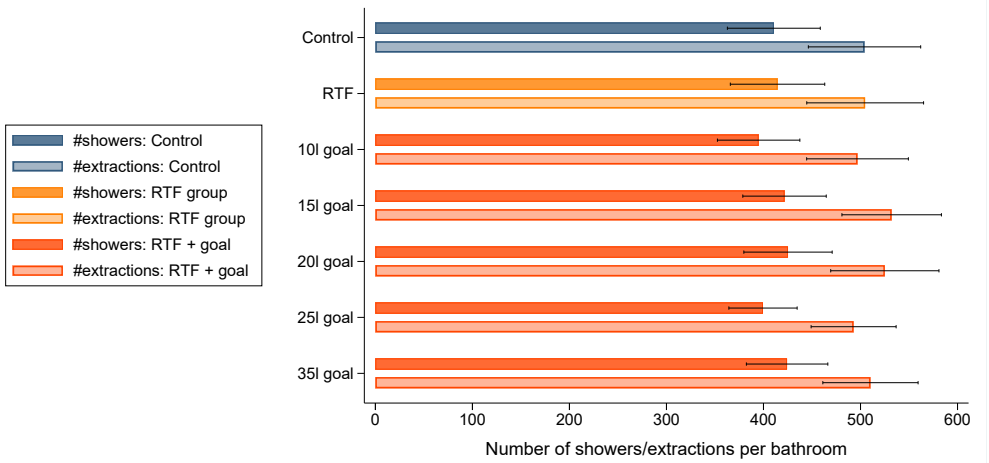


Figure 1.3. Number of showers and water extractions by experimental condition

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

Another issue could be that not all actual showers are recognized as such, because we only record detailed data for water extractions that use at least 4.5 liters. The total number of water extractions per bathroom we observe during the study period is about 510 on average. While the share of non-shower extractions seems relatively large, it should be considered that bathrooms in Singapore are often designed as closed cubicles, and that shower heads are frequently used for cleaning purposes. Still, one may be worried that our treatments had an effect along this margin, for example if individuals become more likely to take longer water flow breaks within showers in a way that a single shower is mistakenly recorded by the device as several extractions instead. Therefore, we additionally compare the total number of all water extractions per bathroom by treatment condition in Figure 1.3. Again, there are no significant differences across groups in our sample ($p = 0.9766$).

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

1.3.5 Baseline water consumption behavior

The baseline period of twenty showers per device at the beginning of the study allows us to gain insight into households' water consumption behavior in the shower in the absence of any real-time feedback or goal-setting interventions. Summary statistics are presented in Appendix Table 1.B.1. The average shower in the baseline period lasted 4.9 minutes (excluding breaks in water flow) and used up about 20.03 liters of water, which is about 50% less compared to earlier studies using the Amphiro

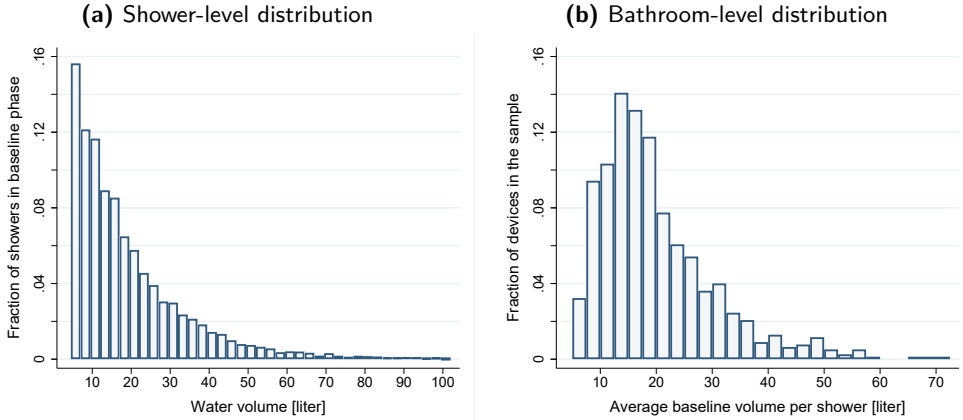


Figure 1.4. Baseline water usage per shower in liters

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

smart meter in non-tropical countries (Tiefenbeck, Goette, et al., 2018; Fang, Goette, Rockenbach, et al., 2020; Byrne, Goette, Martin, et al., 2021). One reason for this is the relatively low flow rate of 4.60 liters per minute on average, which is perhaps partly due to overall lower water pressure in the high-rise HDB buildings, and partly due to the use of instant heaters as opposed to central hot water heating. Another reason is that Singapore's climate is very warm and humid, which often necessitates short showers in the middle of the day to rinse off the sweat and freshen up. This is also reflected in a low average water temperature of 33.8 degrees Celsius and a high shower frequency of 1.3 showers per day.

Figure 1.4a plots the histogram of water volumes based on more than 15,000 showers in the baseline period. The distribution is heavily right-skewed, with a significant share of ultra-short showers (30.5%) that require less than 10 liters of water. The median shower only uses 14.9 liters. However, there is a long tail of showers with significantly higher water consumption, with the 90th percentile lying at about 40 liters. The histogram of average shower volumes at the bathroom-level in Figure 1.4b shows that there is still large heterogeneity in baseline consumption behavior across households and bathrooms, but the distribution becomes more concentrated and less heavily skewed, indicating substantial within-household heterogeneity of showers. Indeed, only 37.6% of the variation in baseline shower volumes is explained by across-bathroom heterogeneity. This can be driven both by differences across individuals who use the same bathroom as well by longer and shorter showers taken by the same individual. Three outlier bathrooms with an average baseline volume of more than 60 liters per shower, which can be spotted at the far end of the histogram, will be excluded for all formal analyses.

Table 1.3. Randomization checks

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

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

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

1.3.6 Randomization checks

Our identification strategy relies on randomization producing treatment groups that are comparable with regard to observable and unobservable subject characteristics. It is naturally impossible to test the latter, but Table 2.4.2 shows good balance based

on a number of key observable variables with regard to baseline behavior. Crucially, average water usage per shower is comparable across the seven experimental conditions, and a joint F-test detects no significant differences overall ($p = 0.937$). This is of particular importance as other studies generally find that households or individuals with high baseline consumption tend to respond more strongly to policy interventions (e.g. Allcott (2011), Ferraro and Price (2013b), and Tiefenbeck, Goette, et al. (2018)). Furthermore, there is no evidence for significant pre-intervention differences along other behavioral margins in the shower, namely duration of the shower, average water flow rate, and water temperature. While there is a single t-test that indicates significantly lower baseline water temperature in the 15L group relative to the Control group at the 5% level, this is in line with the rate of false positives one would expect due to multiple testing, and the F-test cannot reject the null hypothesis of joint equality across all groups ($p = 0.156$).

We further use data from the baseline survey to check for balance with regard to water conservation attitudes as well as general environmental and cost-consciousness attitudes from the baseline survey, because these could determine how individuals respond to our water conservation interventions. Appendix Table 1.B.3 shows that there are no significant differences in these attitudes across groups, further indicating that we can use the randomly assigned treatments to estimate the causal effects of real-time feedback and exogenous goals in our setting.

1.4 The main experimental outcomes

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

1.4.1 Descriptive evidence

In Figure 1.5a, we plot the moving average of water usage per shower over the course of the study. For this purpose, we construct a study progress variable that is coded to take values between 0% (beginning of the study) and 100% (end of 4-months study period).[15] Recall that in all experimental conditions, we included a baseline period of 20 showers per device at the beginning to collect behavioral data in the absence of any intervention. To clearly illustrate changes in water usage when the real-time

15. Study progress of households who received the devices for six months is coded between 0 and 150. For these households, the months 5 and 6 are not presented in Figure 1.5a, as the trends would become very volatile due to the drastic drop in the number of observations. As the shower meter does not store global time, we construct a measure of study progress using the order of showers and assuming constant shower frequency.

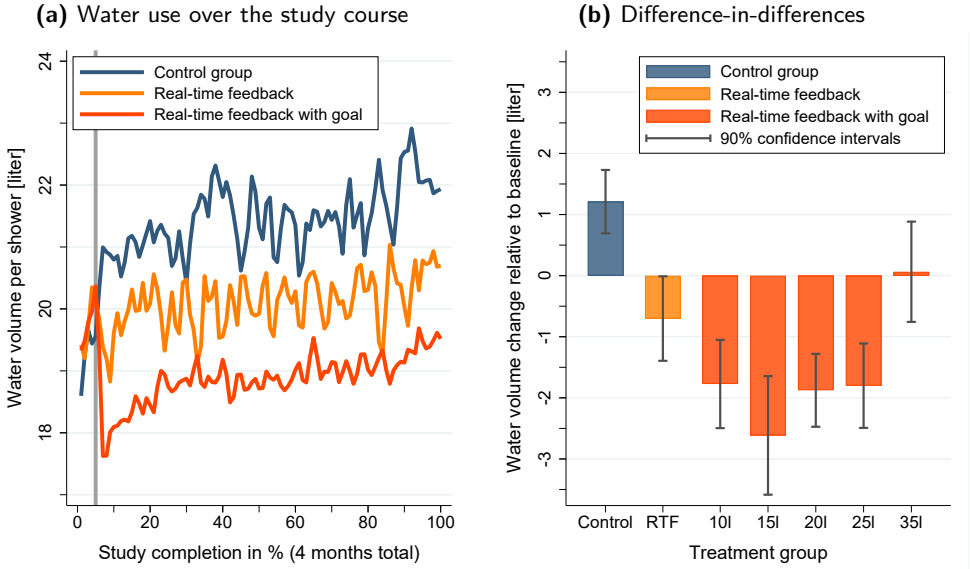


Figure 1.5. Descriptive evidence on the effects of our interventions

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

feedback and goal-setting interventions started in the respective treatment groups, we normalize the baseline period to end at 5% study progress for all households.

The average volume per shower is about 20 liters at baseline, with a slight upward drift that continues over the entire study period in the Control group (blue line). In contrast, we observe a sharp and instant drop in water usage in the RTF and Goal conditions once the intervention started, and the conservation effects remain stable over time, with all lines following close to parallel trend. Subjects who only receive real-time feedback consistently use about 1-2 liters of water less per shower relative to the Control group. The graph also shows that the pooled Goal conditions appear to have a stronger effect than real-time information alone, as the average water volume per shower lies consistently below the outcomes of the RTF condition. It is particularly noteworthy that goals and real-time feedback immediately unfold their full impact from the first shower in which they become active. This suggests that the behavioral responses are driven by higher effort or attention rather than by gradual learning about how to shower more water-efficiently.

In order to get a more accurate sense of the changes induced by the different treatments, we take the average water use for each household during the intervention phase and subtract from it the household's average water use during the base-

line phase. This reduces the number of observations to one per bathroom and allows us to perform a graphical difference-in-difference analysis with valid standard errors. The results are displayed in Figure 1.5b. The leftmost bar in the figure shows the average change in water use per shower during the intervention phase compared to the baseline phase for the control group. As was visible in Figure 1.5a before, there is an upward drift in the Control group of more than one liter per shower on average. By contrast, the RTF group experiences an approximately 0.7 liters decrease in water volume compared to the baseline period. The difference-in-difference estimate of the treatment effect is thus slightly below 2 liters per shower, and the 90% confidence intervals around the two means are far apart from each other, thus suggesting that the difference a strongly statistically significant.

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

1.4.2 Average treatment effects

While the previous analyses in Figure 1.5 already provided descriptive evidence of the effects of real-time feedback and goals, we now exploit the full panel structure of the data to obtain more efficient estimates of the average treatment effects. We do so by estimating the following statistical model:

$$y_{is} = \alpha_i + \beta_R T_{R,is} + \beta_{10L} T_{10L,is} + \ldots + \beta_{35L} T_{35L,is} + \delta_t + \epsilon_{is} \qquad (1.1)$$

where y_{is} is water use in shower s recorded by device i. The coefficient α_i is a device-level fixed effect that is identified through the baseline period of 20 showers at the beginning. $T_{k,is}$ are indicator variables for different treatment groups k and equal 1 if the shower occurred in the intervention phase ($s \geq 21$) and the device i belongs in the respective treatment group. The RTF group is indicated by subscript R and the Goal groups are indicated by their specific volume target (10L, 15L, 20L, 25L, 35L). The control group is omitted and serves as the reference group. Due to random

Table 1.4. Impact of feedback and goals on water consumption per shower

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

Coefficients obtained from estimating equation 1.1. Only includes devices with more than 20 showers in total. Three outliers with average baseline volume of above 60 liters are dropped. Column (2) to (4) only use subperiods split at the 40% and 75% study progress marks. Standard errors in parentheses clustered at household level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

assignment of households into experimental conditions, the coefficients β can be interpreted as the average treatment effects (ATE) of each treated group. We model time fixed effects by a study progress variable discussed previously, captured the coefficient δ_t for percentile t of the study duration. ϵ_{is} is the shower-specific error term. As many showers are observed for the same household on possibly up to two devices, the observations cannot be considered independent within a household. Therefore, we allow for an arbitrary covariance matrix of residuals within households by calculating heteroskedasticity-robust standard errors clustered at the household level (Abadie, Athey, Imbens, et al., 2017).

Table 1.4 column 1 presents the results for the ATEs that come from estimating the difference-in-differences model in equation 1.1. The coefficient estimates closely resemble those from Figure 1.5b. We find that real-time feedback alone al-

ready significantly reduced water consumption by about 1.87 liters ($p < 0.001$) per shower compared to the Control condition that did not receive any feedback. This corresponds to about 9% of the baseline average, which is consistent with previous studies using the Amphiro smart meter when taking into account the baseline differences (e.g. Tiefenbeck, Goette, et al., 2018; Fang et al., 2020). Crucially, we find that exogenously assigned goals can induce conservation effects above and beyond that of real-time feedback alone. For instance, the 15L goal condition reduced water usage by 3.92 liters per shower and thus by significantly more than the RTF condition ($p = 0.003$), with the estimated ATE being about twice as large. However, not all goals are created equal. While the 15L goal was the most effective goal based on the point estimates, the 10L, 20L, and 25L goal all induced a water conservation effect of around 3.0 liters and thus still performed 60% better than real-time feedback without any goal.[16] By contrast, the exceedingly easy 35L goal has does not lead to a stronger conservation effect than the RTF condition ($p = 0.217$), with the point estimate of -1.1 liters indicating that, if anything, it is actually less effective than having no goal assigned at all.

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

1.4.3 Stability of treatment effects over time

The previous results show that, on average, suitable goals can have a strong additional effect on water conservation behavior when added to real-time feedback. Figure 1.5a further indicates that the effects are stable over time when pooling all five Goal conditions. However, it is conceivable that time trends vary depending on the difficulty of the goal (Goette, Han, and Lim, 2021). In order to examine the stability over time more formally, we split the intervention phase in three roughly equally long periods (of about 6 weeks length) and estimate the treatment effects separately

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

for these periods. Columns 2 to 4 in table 1.4 indicate a remarkable stability of effect sizes over the entire duration of the study. While the estimated coefficients exhibit some minor fluctuations over the course of several months, these differences are statistically insignificant for all treatment groups and quantitatively small, well within the range of one standard error. There is also no monotonic pattern that could indicate a clear time trend. At most, the average conservation effect of goals in our study decreases by a magnitude in the order of 0.1 to 0.3 liters per shower from the first weeks to the final weeks of the intervention.

Appendix table 1.B.4 further shows that these results are confirmed when interacting treatment effects with a four-part spline of intervention progress, so the coefficients can be interpreted as the speed with which the treatment effect changes with study progress. Two important conclusions emerge from the analyses here. First, all our experimental treatments have an immediate effect on behavior: literally starting from the first shower of the intervention phase, the treatments are fully effective. Second, the treatment effects remain stable over our intervention period of four to six months. Therefore, there is no evidence that real-time feedback and exogenously assigned goals begin to lose their effectiveness on average conservation behavior as long as they remain in place.

1.4.4 Interaction with baseline usage

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

We analyze heterogeneity by baseline consumption first by splitting the sample into consumers with average baseline water usage that is either above or below the sample median (17.4 liters), and then estimating equation 1.1 separately for these

two subgroups. Second, we also estimate an interacted model

$$\begin{aligned} y_{is} &= \alpha_i + \beta_R T_{R,is} + \beta_{10L} T_{10L,is} + \ldots + \beta_{35L} T_{35L,is} + \gamma_C \times z_{it} \\ &\quad + \gamma_{RT} T_{RT,is} \times z_{it} + \ldots + \gamma_{35L} T_{35L,is} \times z_{it} + \delta_t + \epsilon_{is} \end{aligned} \tag{1.2}$$

where the treatment indicators are interacted with baseline consumption z_i, i.e. average water use during the baseline phase for each household. Notice that even though we have fixed effects in place, we need to allow for a main effect interacting the intervention indicator with z_i, because there could be differential trends associated with different values of z_i, for example due to mean reversion or other sources of baseline-dependent serial correlation. We code z_i to be equal to 0 in the baseline period of the first 20 showers, so these trends will be captured by the coefficient γ_C. The coefficients $\gamma_{RT}, \gamma_{10L}, ..., \gamma_{35L}$ thus indicate by how much the treatment effect changes with a 1 liter increase in average baseline water usage per shower relative to the Control group.

The results are displayed in Table 1.5. Columns (1) and (2) show the estimated treatment effects for below-median and above-median consumers, respectively, and column (3) shows the estimated interaction effects in the linear interactions model from equation 1.2. Consistent with previous literature, we observe that conservation effects are significantly stronger for subjects with high baseline consumption. Real-time feedback alone had no significant effect for low-baseline consumers, who used only 12.49 liters per shower on average in the baseline phase, whereas it reduced water use per shower by 3.25 liters on average for high-baseline consumers, who used 27.18 liters per shower on average in the baseline phase. This is also reflected in an estimated linear interaction of 0.235 liters lower consumption per 1 liter increase in baseline consumption in the RTF group. Note that the relative average treatment effect of real-time feedback was around 9%, hence higher baseline consumption is associated with an overproportional increase in effectiveness, as found in several previous studies of resource conservation (see, e.g., Allcott, 2011; Allcott and Rogers, 2014; Tiefenbeck, Goette, et al., 2018).

The treatment effects for the goal conditions exhibit qualitatively similar interactions with baseline consumption, but there is also significant variation in the extent of heterogeneity induced by different goal difficulty level. Indeed, we can rule out at the 1% level that the interaction effects in column (3) are equal among all five goal conditions ($p = 0.0084$). Column (1) shows that even in the subsample of low-baseline consumers, where real-time feedback alone was ineffective, all goal conditions except for the 35L group induced statistically significant conservation effect of 1.59 to 2.17 liters per shower, which is equivalent to 13% to 17% of baseline consumption. Although we cannot reject the null hypothesis of equal effects in the four goal conditions ($p = 0.651$), it is worth noting that the point estimate is largest for the most difficult 10L goal, which achieved a reduction in water consumption by 2.17 liters, which is significantly more than in the RTF condition ($p = 0.0032$).

Table 1.5. Heterogeneous effects by baseline water consumption

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

Columns (1) and (2) estimate equation 1.1 separately for devices with below- and above-median baseline consumption. Column (3) shows the coefficients for interaction effects from estimating equation 1.2. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

However, despite its impressive performance among low-baseline users, the effect of the 10L condition in the subsample of high-baseline users (-3.62 liters) was comparable to that of the RTF condition ($p = 0.7140$) — accordingly, its linear interaction coefficient in column (3) is also closer to zero ($\beta_{10L} - \beta_R = .1131, p = 0.058$). This finding is consistent with the theoretical prediction that a goal so difficult that it becomes unattainable does not have strong effects. As baseline consumption of a household increases, attaining the 10L goal becomes subjectively harder, thus its additional motivational power eventually vanishes. On the other extreme, the easy 35L goal had no significant conservation effect for low-baseline users, likely because it

is not challenging enough and thus simply ignored.[17] Perhaps more surprisingly, the 35L condition was in fact less effective than the RTF condition for high-baseline users ($\beta_{35} - \beta_R = 1.826, p = 0.0723$), which may be suggestive evidence for boomerang effects or crowding out of intrinsic motivation to reduce water consumption in response to feedback. As a consequence, the interaction with baseline consumption is very low and insignificant.

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

In Figure 1.6, we further illustrate the relationship between behavioral responses and baseline consumption in a nonparametric way by estimating local linear regressions at the bathroom level for each experimental condition separately. Note that we cut off the graph to the right, because the confidence bands for devices with the highest baseline usage become very wide. The local linear estimates confirm the results in Table 1.5 that real-time feedback is mostly ineffective for consumers who were already very water-efficient, but starts to become effective for households with an average baseline usage of above 15-20 liters, with the water conservation effect now increasing approximately linearly compared to the control group. For the Goal conditions, the pattern is in principle similar, but varies across difficulty levels. In the 10L and 15L goal conditions, even households with low baseline usage of around 10 liters per shower already show relatively large conservation effects, but the effect estimates converge to those of the RTF condition for high-baseline consumers, as the goals become too challenging. Indeed, the slope is almost generally flatter in the 10L condition compared to the RTF condition. In contrast, the estimates for the 15L condition exhibit a steeper slope in the range between 10 liters and 25 liters baseline usage, which is where the majority of households fall into (see Figure 1.4b).

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

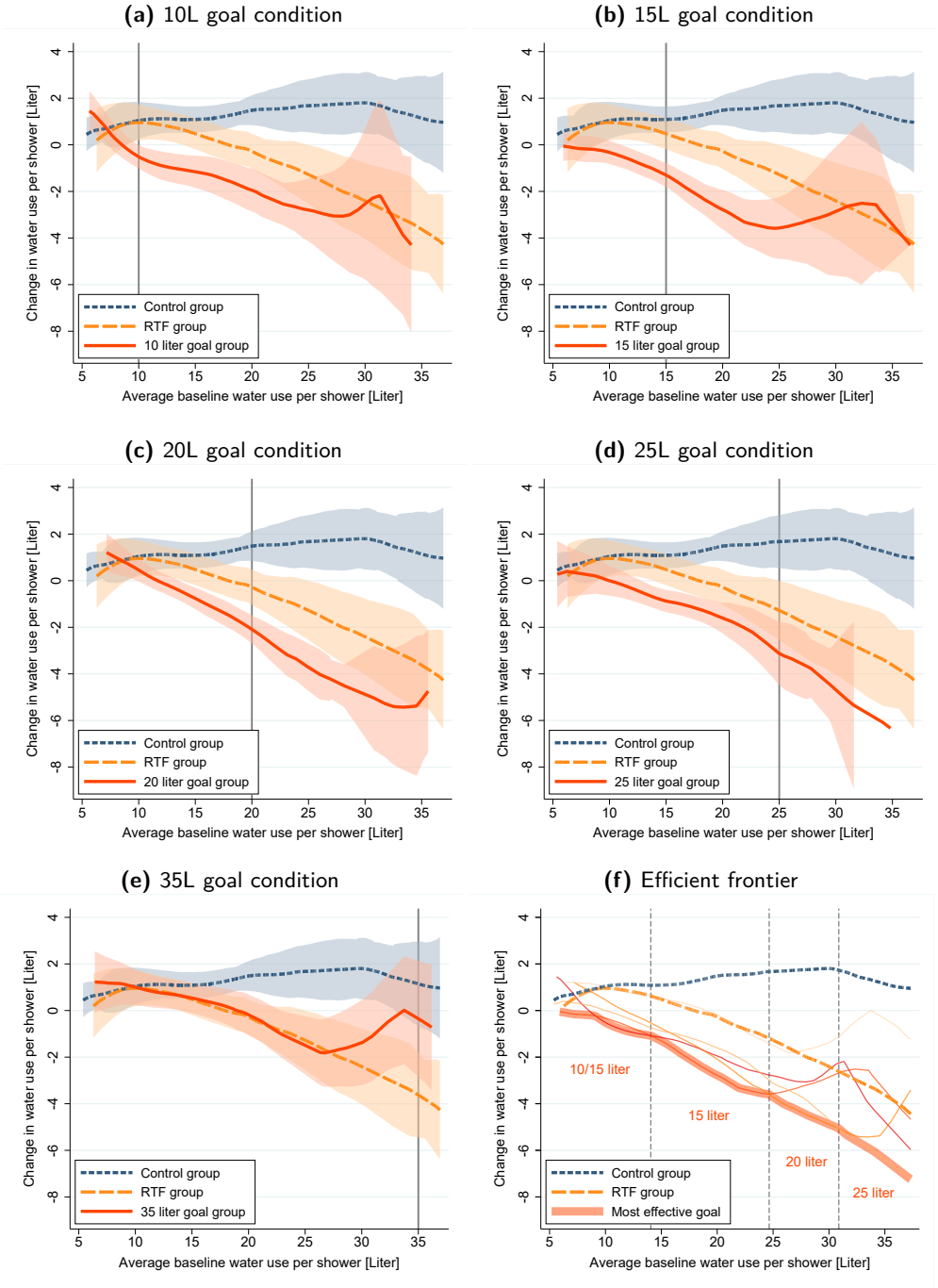


Figure 1.6. Local linear regressions of DiD estimates by baseline consumption

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

The estimates for the 20L and 25L goal conditions roughly resemble the estimates for the RTF group with a downward parallel shift, whereas the local effects in the 35L goal group are almost identical to the RTF group except for high baseline users, for which real-time feedback without goals is actually more effective.

Figure 1.6f compiles the nonparametric fits for all treatment groups in a single graph, which allows us to trace out the treatment effect "frontier" based on the most effective goal (based on the point estimates in our sample) as a function of baseline consumption. A highly suggestive pattern arises: at the lower end of the baseline distribution, the 10L and 15L goal conditions induce the largest conservation effects; in the middle of the distribution, where the largest share of households fall into, the 15L goal performs best; at the higher end of the distribution, the 20L condition and 25L condition start surpassing it. This pattern again supports the notion that moderately challenging goals are most motivation-enhancing, where the optimal goal may vary across individuals due to differences in subjective difficulty levels. However, the easy 35L goal breaks with the pattern to a certain degree, as it seems to become counterproductive exactly for the subset of households for whom achieving it is not a sure-fire endeavor anymore. This could be explained in a way that externally set goals also represent a type of socially acceptable standard, which may crowd out potentially more ambitious personal standards.

1.5 Behavioral mechanisms of goal-setting

1.5.1 Excess mass at the goals

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

We do so by first exploiting the random assignment into experimental conditions to compare the empirical distributions of showers in the intervention phase between the goal groups and the RTF group. If conservation goals serve as reference points for evaluating success and failure, e.g. by creating a kink (loss aversion) or a notch (fixed reward) in the utility function, then we would expect a general shift in probability mass from above the goal to below the goal, and specifically also bunching of outcomes at the respective goal (Kleven, 2016). For example, Allen et al. (2017) provide evidence that the distribution of marathon runners' finish times exhibits excess mass below and missing mass above round numbers (e.g. 3h, 3:30h, ...).

The advantage of our setting is that we have experimentally-induced variation in both whether households receive a goal at all and what the specific goal is, and thus do not need to rely on smoothness and local boundedness assumptions to construct a

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

In Figure 1.7, we group shower observations during the intervention period into 1 liter bins based on their distance to the respective goal and plot the excess and missing mass of showers in goal group versus RTF group households. The visual impression is striking. Assignment of an exogenous conservation goal induces a consistent shift in probability mass from above to below the goal, thus providing compelling evidence that individuals exert effort in order to avoid exceeding the target level that was externally assigned to them. Moreover, the shifts in the empirical density function are not uniform. There is strong bunching in the 1 liter bin just before the respective goal, with showers in the goal conditions have a $0.68\%p$ higher probability to fall into this bin, which corresponds to a relative increase in 25% compared the respective share of showers in the RTF condition (2.7%).[18] The spike in distribution just before the goal is followed by a sharp drop in the relative share of showers just above the goal, although still remaining slightly higher than in the absence of an explicit goal. The largest amount of missing mass is found at about 5 to 10 liters above the goal, after which the distributions converge again at a slow rate. While bunching is most evident just below the goal, there is an excess mass of showers up to 20 liters below the goal relative to the RTF condition, which suggests that the influence of goal-setting on consumption behavior is not limited to extremely local responses around the goal. Note that due to the water volume of a shower being bounded from below by 4.5 liters, each goal condition is only represented from $-G + 4.5$ onwards in Figure 1.7, where G is the conservation target. Appendix Figure 1.A.3 compares the distribution of each goal condition separately with the distribution in the RTF group. Interestingly, we observe missing mass of ultra-short showers in the 35L group, which again suggests a boomerang effect for very easy goals.

To estimate local bunching around conservation goals more formally, we estimate a linear probabilities model with an indicator for a shower falling in a partic-

18. The distribution of showers with regard to goal distance are presented in Appendix 3.A Figure 1.A.2.

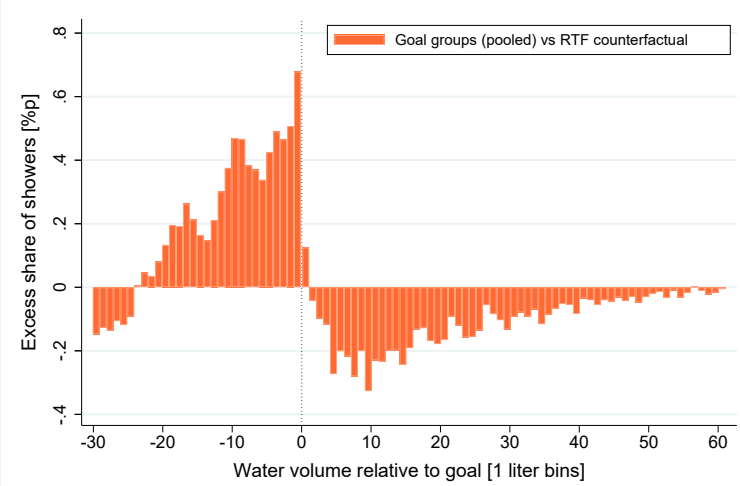


Figure 1.7. Excess share of showers in the goal groups relative to the RTF group

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

ular volume bin Δ_V close to a salient thresholds V (e.g. 10 liters, 15 liters, ...) as dependent variable:

$$\mathbb{1}\{y_{is} \in \Delta_V\} = \alpha_i + \beta_1 T_{is} + \left(\beta_2 + \gamma\, \mathbb{1}_i\{V = G\}\right) T_{is}^{goal} + \delta_s + \theta_V + \epsilon_{is}\,. \quad (1.3)$$

We include fixed effects for bathroom (α_i), intervention period (δ_s), and threshold (θ_V). T_{is} is an indicator for a shower by a treated household (both RTF and goal groups) in the intervention period, and T_{is}^{goal} is an indicator for intervention period showers by households in one of the five goal conditions. We additionally interact the latter with a match indicator that takes the value 1 only if the threshold coincides with the conservation goal. The coefficients β_1 and β_2 indicate whether households who receive real-time feedback use salient numbers as anchor that do not correspond to an externally-assigned conservation target. The main coefficient of interest γ captures how much more likely it is that showers fall in a certain bin close to the goal relative to other salient thresholds. Thus, we exploit the random assignment of different goals to households to identify changes in the local distribution around a goal.

We consider three bin sizes $|\Delta_V| \in \{0.5L, 1L, 2L\}$ and estimate equation 1.3 separately for these bins above and below the thresholds. Table 1.6 presents the results of this empirical exercise. Columns (1) to (3) show that there is significant bunching of showers at the goal. For example, showers were 3.1 percentage points (19%) more likely to be placed less than 1 liter below a threshold that corresponds to an exogenous goal, and 2.6 percentage points (27%) more likely to be placed less than half a liter below a goal. On the other hand, there are only quantitatively weak signs of missing mass up to 1 liter above a goal threshold, but the share of showers that are up to 2 liters above a goal is 2.5 percentage points (9%) lower. In contrast to

Table 1.6. Probability of showers just above or below a salient threshold

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

Notes. Results come from estimating equation 1.3 using ordinary least squares. The dependent variable is an indicator for whether a shower falls into a particular volume bin around a salient threshold. We consider thresholds in steps of 5 from 10 liters to 45 liters. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Allen et al. (2017), we find no evidence of strong local responses to salient thresholds that are not associated with an explicit external goal. However, this does not necessarily imply that individuals who receive real-time feedback do not attempt to set and achieve personal conservation goals, as discontinuities in the distribution may as well be hidden by heterogeneity in self-set goals.

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

In general, however, inferring local behavioral responses from excess mass in the empirical probability density functions can be partly complicated due to broader shifts in the cumulative density of water consumption levels in response to feedback and goals. It can thus be hard to interpret excess mass in a certain range, as it could be driven both by a local change in the probability of stopping a shower or a general shift of high-volume showers to showers with lower volume. Therefore, in the next step, we examine the stopping probabilities of individual shower in terms of the

hazard rate, i.e. the probability that a shower stops at a given water consumption level *conditional* on "surviving" until this point.

1.5.2 Goal distance and stopping hazards

To give a graphical overview of how goals affect the stopping probabilities of showers in Figure 1.8, we again pool all five goal conditions and calculate the hazard rate as a function of the distance to the goal in steps of deciliters, our most fine-grained unit of measurement. The hazard rate at point k is defined as the conditional probability of stopping between $k-1$ and k deciliters relative to the goal, given that the relative water volume is above $k-1$ deciliters. Hence, a higher hazard rate reflects a higher probability to end the shower at a given point and thus higher effort to conserve water, irrespective of where k lies in the distribution. As before, we construct the counterfactual hazard rate for the goal groups by assigning placebo goals to each observation in the RTF group five times. To flexibly control for baseline differences across experimental conditions, we further adjust the hazard rates in the intervention period by dividing through local linear estimates of the baseline hazard ratio between the goal groups and the RTF group.[19] Thus, the following results can be interpreted as difference-in-differences of hazard rates.

Figure 1.8a plots the hazard rates as a function of water volume relative to the conservation goal in deciliters, as well as smoothed estimates using local linear regressions. In addition, Figure 1.8b plots the hazard ratio relative to the RTF counterfactual using the smoothed hazard rate estimates, with pointwise confidence intervals obtained from a block bootstrapping procedure that accounts for clustering at household level.[20] The counterfactual hazard rate stays relatively constant, fluctuating around 1.25% with a slight downward trend. Some wave-like patterns with humps at round numbers hint at the presence of self-set goals à la Allen et al. (2017), but are too small to be detected in Table 1.6. In comparison, the hazard rates in the goal conditions show a very clear pattern. Stopping behavior is relatively unaffected by the exogenously assigned goal when the water volume is still more than 15 liters below the goal, as there is large remaining scope for finishing the shower in time.[21]

19. More specifically, we run separate local linear regressions of the baseline hazard rates by goal distance for the goal conditions and the RTF condition with Placebo goals. We then use the smoothed estimates to calculate the local hazard ratios and divide the intervention period hazard rates in the goal conditions by the respective hazard ratio. The results without any adjustment for baseline differences can be found in Appendix Figure 1.A.4 and look very similar.

20. Specifically, we resample our data $4,000$ times by drawing household (and their entire time series of data) with replacement and estimating the non-parametric regressions for each bootstrap simulation and then constructing equal-tailed percentile confidence intervals. Note that we intentionally undersmooth the local linear hazard rate estimates for statistical inference to ensure that the bias term shrinks faster than the variance term.

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

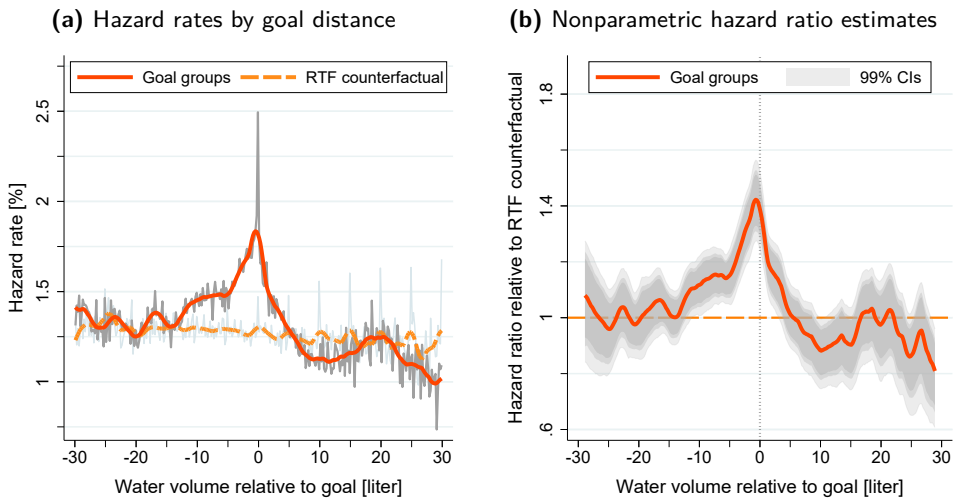


Figure 1.8. Stopping hazard by goal distance

Notes: The left panel (a) plots the hazard rates of showers by deciliter of distance to the respective conservation goal. Smoothed estimates are obtained through local linear regressions using the Epanechnikov kernel with bandwidth 0.6 liters. Hazard rates are adjusted for baseline differences between experimental conditions by dividing through smoothed local estimates of the hazard ratio between goal and RTF condition in the baseline period (see footnote 19). The right panel (b) plots the hazard ratio, calculated from the ratio of smoothed hazard rate estimates in the goal group and the RTF counterfactual. Bootstrap percentile confidence intervals are obtained from clustered bootstrapping with 4,000 simulations, using households as unit of resampling. Different shades of grey reflect 90%, 95%, and 99% confidence regions, respectively.

As individuals approach the goal, the hazard rate increases above the counterfactual rate and reaches its peak just below the goal. What springs to the eye is the enormous spike in the stopping probability at the very last deciliter, which jumps from about 1.75% up to 2.5% and then immediately down again. While the smoothed estimates generally track the movements of the empirical hazard rate of the goal conditions very well — capturing about 84% of the variation within 30 liters around the goal — they fail to account for the anomalous spike at the goal.[22] This is perhaps the single most powerful piece of evidence in this study that individuals respond to non-binding, exogenously-assigned goals. Interestingly, the hazard ratio rapidly reverts and becomes statistically indistinguishable from 1 after just three to four liters since the goal has been missed, even dropping below 1 for showers with higher water

22. We can quantify the anomaly by fitting a local linear estimate that uses all empirical hazard rates except for the one at the last deciliter before the goal, in the spirit of the bunching estimator approach by (Chetty, Friedman, Olsen, et al., 2011). Comparing the actual hazard rate to the leave-one-out estimate indicates a discontinuous jump by 0.76 percentage points at the goal, which corresponds to about 44%. Using clustered Monte Carlo bootstrap inference, we can show that this jump is highly statistically significant, as in 4,000 bootstrap simulations there was not a single instance in which no large positive spike in the hazard rate occurred.

volumes, which stands in contradiction to loss aversion models, which would predict *higher* stopping rates in the loss domain, i.e. when the goal has been missed, compared to the gain domain.

This setup allows us to test the predictions of the loss-aversion and fixed-penalty model from a different angle. If loss aversion is driving goal effects, then quitting hazards should be unaffected (up to some uncertainty owing to randomness in stopping) before water usage has reached the goal. The stopping hazard should increase once the individual is past the goal and in the loss domain with the correspondingly higher marginal disutility from water use. By contrast, the fixed-penalty model implies that stopping hazards should be higher as the individual approaches the goal. Since the penalty is fixed and incurred as the individual surpasses the goal, the individual has an incentive to stop somewhat early owing to the randomness in the water used.[23]

This pattern is fully consistent with the fixed-penalty model: individuals stop somewhat ahead of the goal in order to avoid overshooting due to randomness. However, once they overshot, goal-related efforts to stop vanish and the stopping hazards becomes indistinguishable from those by individuals who received real-time feedback without any explicit goal. At the same time, the pattern is difficult to reconcile with the loss-aversion model, in which the higher marginal disutility from surpassing the goal motivates stopping efforts, as the stopping hazard in the goal conditions quickly reverts to the one of the RTF group once the goal is missed.

1.5.3 Changes in behavioral response over time?

The underlying behavioral mechanism of how goals enter the utility function also has implications of the stability of the treatment effects over time. If one takes the view that goals take on the role of reference points directly (See, e.g., Heath, Larrick, and Wu, 1999), then responses should remain stable over time. However, in a model of expectation-based reference points Koszegi and Rabin (2006) and Koszegi and Rabin (2009), it is possible that goals may not only affect reference points directly, but also shift expectations.[24] In such a model, a shift in expectations can be self-fulfilling and subsequently affect behavior. However, this raises the question of whether the impact of goals becomes less effective over time. Suppose an individual was assigned a hard goal (compared to her baseline water use). If this affects her expectations and thus her reference point, both of the models outlined above would predict an increased conservation effort. However, as time goes by and the individ-

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

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

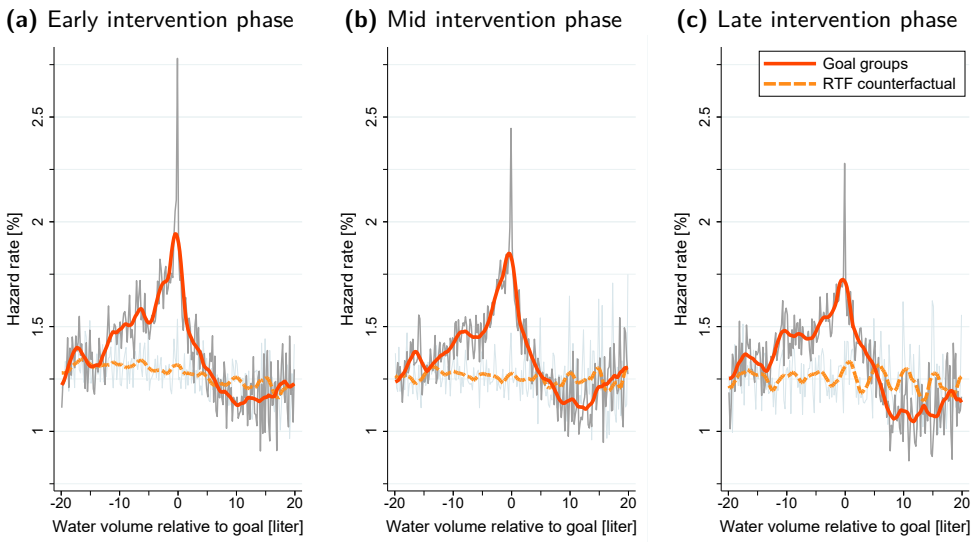


Figure 1.9. Stopping hazards over time

Notes: Hazard rates of showers by deciliter of distance to the respective conservation goal, split by three phases of the intervention period. Smoothed estimates are obtained through local linear regressions. All procedures follow the ones in Figure 1.8a.

ual repeatedly falls short of the goal, this may affect her expectations, and thus her reference point. Thus, it is possible that goal effects are temporary and gradually losing their effect on behavior.

In Figure 1.9, we further split the data into the three phases of the intervention period to examine whether behavioral responses adjust over the course of several months. The first observation is that stopping hazards for individuals who only received real-time feedback remain fairly stable, mirroring the results for average water consumption from Table 1.4. The second observation is that, qualitatively, the pattern induced by exogenous goals also remains similar in the later phases of the intervention, with stopping hazards gradually increasing starting from 10 to 15 liters below the goal, peaking with an anomalous spike at the goal, and then quickly plummeting again. However, the third observation is that, quantitatively, the peak at the goal diminishes considerably in magnitude over time. In the first weeks of the intervention, the hazard rates exhibits an impressive jump by 53% (0.96 percentage points) to 2.76% at the goal, whereas in the final weeks it "only" goes up by 38% (0.63 percentage points) to 2.26%. We corroborate this finding in Appendix Table 1.B.5, which extends the analysis in Table 1.6 by an interaction with study progress and shows that bunching of showers in the 0.5 liter and 1 liter bins (but not the 2 liter bin) below a goal decreases significantly over time, with point estimates implying that the excess mass vanishes completely after approximately 6 months.

While we have shown previously in section 1.4.3 that the average water conservation effects induced by nonbinding goals remain largely stable over the duration of our study, our data paints a more nuanced picture when also considering the

Table 1.7. Overall goal attainment rates over time

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

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

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

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

The results are displayed in Table 1.7. In the baseline period, about 62% of showers would have met the conservation goal when pooling all difficulty levels. Columns (1) and (2) show that, hypothetically, attainment rate would have been higher for the RTF condition in the intervention period, which is simply due to the conservation effects in response to real-time feedback shown previously. As there is a slight general upward trend in water consumption levels in the months of our study (see Figure 1.5a), we observe corresponding decreases in hypothetical attainment rates from the beginning to the end of the intervention phase by about 1%p in the Control group and 1.5 %p in the RTF group, both statistically insignificant. When looking at actual attainment rates in the Goal conditions, column (3) shows that after an initial jump by 8% at the start of the intervention, the effect actually decreases by $3.8\%p$ by the end of the four months study duration. This decrease is significantly larger than in the Control placebo ($p = 0.008$) and the RTF placebo ($p = 0.064$), suggesting that it is not only driven by broader trends in water consumption levels but also by goal-specific behavioral mechanisms. To further verify this, we additionally control for water volume fixed effects in column (4). If there was only one goal level in the sample, water volume would perfectly explain goal attainment in this specification. Thus, any non-zero coefficients can only be due to variation in the share of showers below a certain goal threshold relative to households who were assigned a different goal, e.g. when the likelihood of showers below 15 liters increases overproportionally in the 15L goal group.[25] The estimates show that, even conditional on water volume, goal success becomes significantly more likely once the intervention begins, again demonstrating that individuals respond specifically to the goal that was randomly assigned to them. Crucially, the conditional goal attainment effect drops by more than 50% by the end of the 4-month study period ($p = 0.004$).

Overall, the evidence here suggests that over the course of several weeks and months, individuals respond less to the goals that were assigned to them at the beginning of the study, as they gain a more nonchalant perspective on the feasibility and importance of missing or achieving that particular conservation goal. Nevertheless, we see in Section 1.4.3 that there is no evidence for a significant decrease in average water conservation over the 4-month study period. Hence, externally set goals seem to retain a status as vague norm or default about water consumption levels even though the precise target numbers associated with them become less psychologically binding.

1.6 Concluding remarks

In this paper, we presented evidence from a randomized field experiment in the context of household water conservation to examine the effectiveness of goal set-

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

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

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

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

waning local effects, may be resolved by individuals forming habits or setting personal targets that replace the externally set goal.

Overall, our study suggests that goal-setting (and real-time feedback) have the potential to be integrated into simple and easily scalable interventions to encourage desirable behavioral change for example in the domain of pro-environmental behavior, as modern digital technologies are becoming ever cheaper and more advanced.[26] Future research may also consider the comparison between the effectiveness of self-set goals and externally set goals such as the ones we use in this study. Another important question is whether the effects of our interventions are limited only to that targeted activity, showering, or whether there are spillover effects to other water-consuming activities in the household. In a companion paper, we utilize billing data of households that participated in our experiment and observe statistically significant conservation effects of our interventions also in overall household water usage (Schmitt, Tiefenbeck, Fang, et al., 2021). Interestingly, the point estimates suggest quantitatively large *positive* spillover effects, i.e. reductions in water usage also outside of the shower, although we lack statistical power to detect spillover effects more precisely. An interesting avenue for further research is whether different types of interventions have different effects on the sign and size of spillover effects to non-targeted activities, as this may have important implications for cost-benefit calculations.

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

Appendix 1.A Supplementary figures

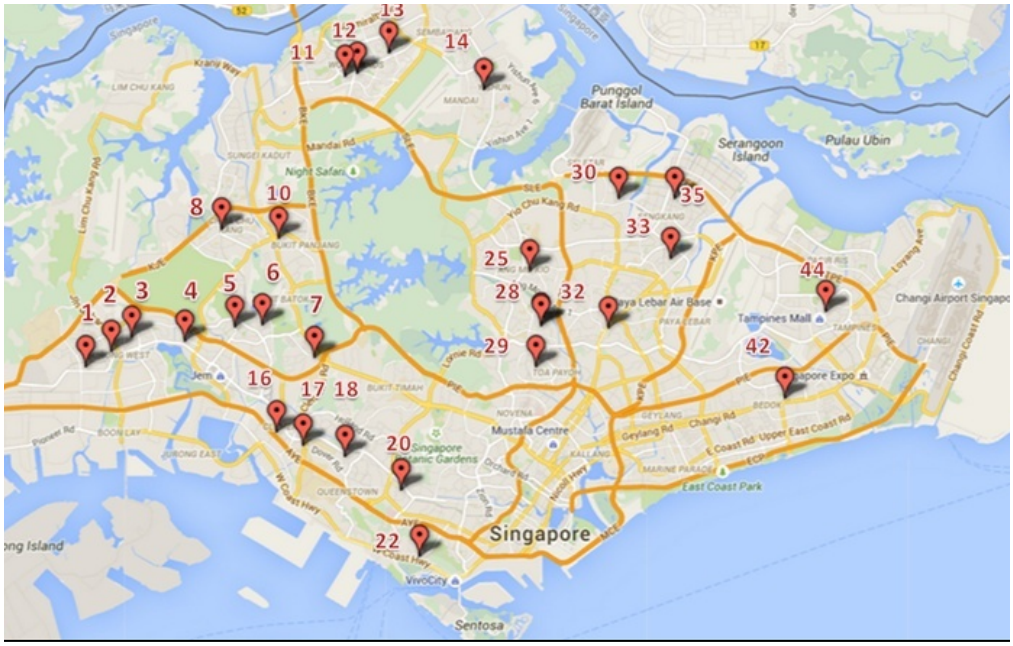


Figure 1.A.1. Sites of participating households in Singapore.

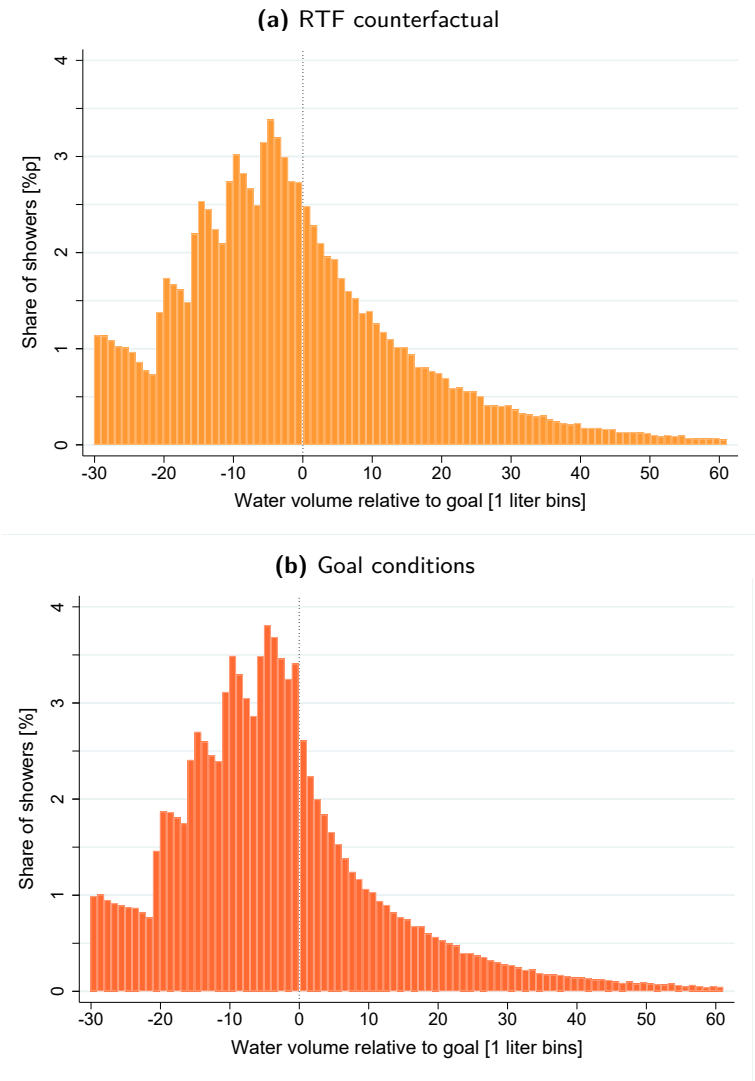


Figure 1.A.2. Distribution of intervention period showers

Notes: Distribution of showers by water volume relative the goal. To construct the counterfactual distribution, we assign placebo goals to households in the RTF condition, where we duplicate each observation there 5 times to assign each possible goal difficulty level once.

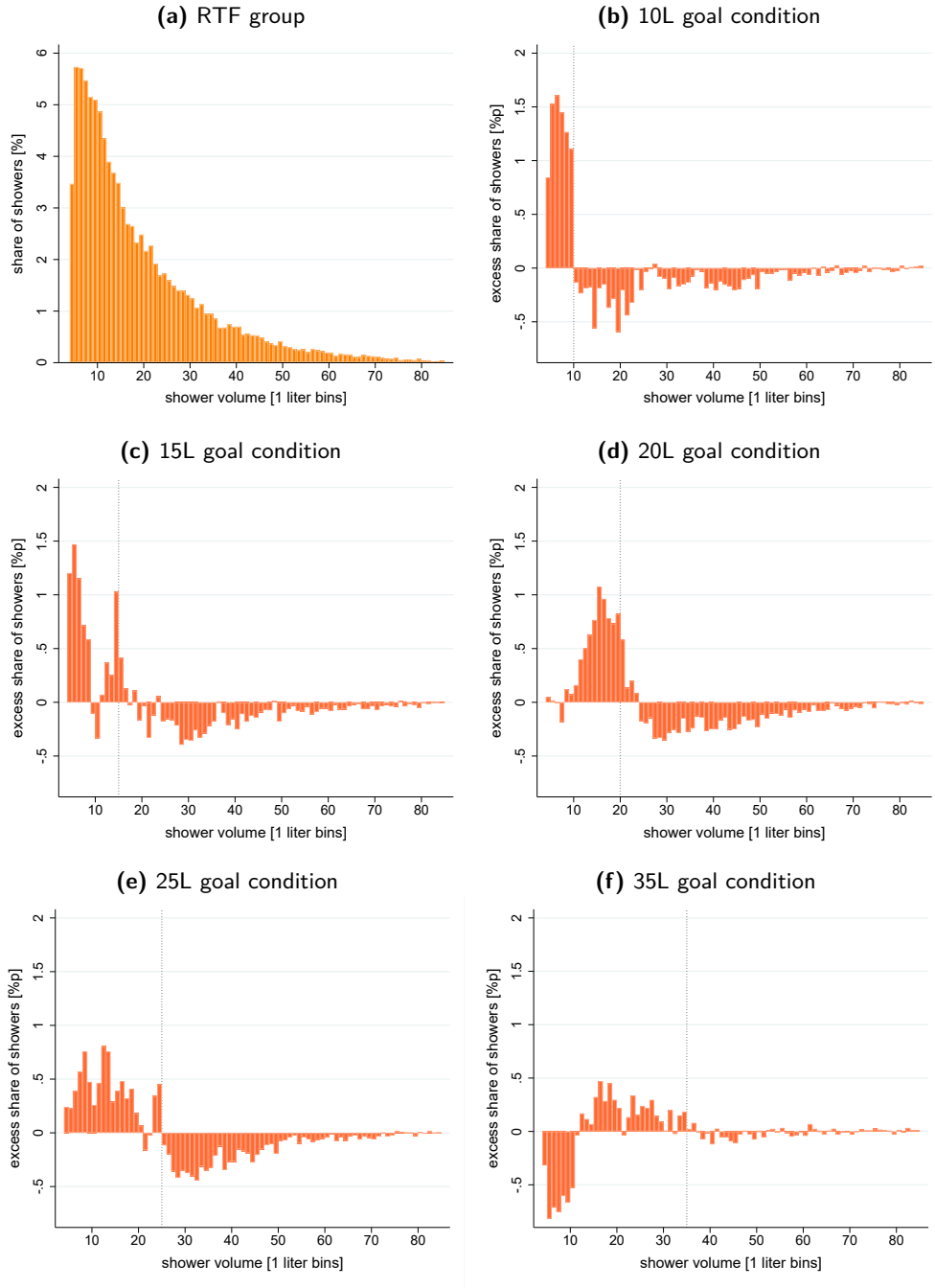


Figure 1.A.3. Excess mass of goal groups relative to RTF group

Notes: Figure (a) plots the distribution of shower volume in the RTF group during the intervention period. Figures (b) to (f) plot the difference in the share of showers in a particular volume bin relative to the RTF group.

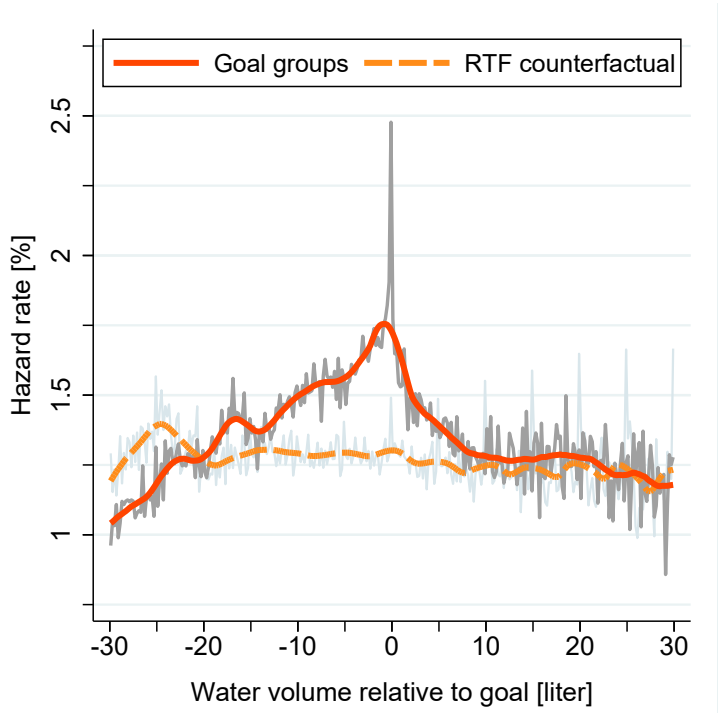


Figure 1.A.4. Stopping hazard by goal distance — not adjusted for baseline differences.

Notes: Recreates Figure 1.8a using the hazard rates in the intervention period without adjusting for baseline difference between experimental conditions.

Appendix 1.B Supplementary tables

Table 1.B.1. Baseline shower information – overview

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

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

Table 1.B.2. Treatment effect on number of showers

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

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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.B.3. Randomization checks — water conservation attitudes

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

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

Table 1.B.4. Stability of treatment effects: four-part splines with study progress

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

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

Table 1.B.5. Probability of showers just above or below a salient threshold

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

Notes. Results come from estimating equation 1.3 using ordinary least squares. The dependent variable is an indicator for whether a shower falls into a particular volume bin around a salient threshold. We consider thresholds in steps of 5 from 10 liters to 45 liters. Standard errors in parentheses are clustered at the household level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" *NBER Working Paper 24003*, [25]

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision." *American Economic Review* 101 (2): 1–25. [38]

Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter. 2005. "A Review of Intervention Studies Aimed at Household Energy Conservation." *Journal of Environmental Psychology* 25 (3): 273–291. [10]

Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter. 2007. "The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents." *Journal of Environmental Psychology* 27 (4): 265–276. [10]

Allcott, Hunt. 2011. "Social norms and energy conservation." *Journal of Public Economics* 95 (9–10): 1082–1095. [10, 22, 27, 28]

Allcott, Hunt, and Sendhil Mullainathan. 2010. "Behavior and Energy Policy." *Science* 327 (5970): 1204–1205. [9]

Allcott, Hunt, and Todd Rogers. 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10): 3003–3037. [10, 28]

Allen, Eric J., Patricia M. Dechow, Devin G. Pope, and George Wu. 2017. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science* 63 (6): 1657–1672. [6–8, 10, 11, 32, 35, 36, 42]

Andor, Mark A., and Katja M. Fels. 2018. "Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects." *Ecological Economics* 148: 178–210. [10]

Angrist, Joshua D., Sydnee Caldwell, and Jonathan V. Hall. 2021. "Uber versus Taxi: A Driver's Eye View." *American Economic Journal: Applied Economics* 13 (3): 272–308. [42]

Avery, Mallory, Osea Giuntella, and Peiran Jiao. 2019. "Why Don't We Sleep Enough? A Field Experiment Among College Students." *IZA Discussion Paper No. 12772*, [5]

Ayres, I., S. Raseman, and A. Shih. 2013. "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage." *Journal of Law, Economics, and Organization* 29 (5): 992–1022. [10]

Blanes i Vidal, Jordi, and Mareike Nossol. 2011. "Tournaments Without Prizes: Evidence from Personnel Records." *Management Science* 57 (10): 1721–1736. [6]

Brookins, Philip, Sebastian J. Goerg, and Sebastian Kube. 2017. "Self-chosen goals, incentives, and effort." *Working Paper*, [10]

Byrne, David, Lorenz Goette, Leslie A. Martin, Lucy Delahey, Alana Jones, Amy Miles, Samuel Schoeb, Thorsten Staake, and Verena Tiefenbeck. 2021. "The Habit-Forming Effects of Feedback: Evidence From a Large-Scale Field Experiment." *CRC TR 224 Discussion Paper No. 285*, [20, 40]

Camerer, Colin, Linda Babcock, George Loewenstein, and George Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *Quarterly Journal of Economics* 112 (2): 407–441. [11]

Carlsson, Fredrik, Christina Annette Gravert, Verena Kurz, and Olof Johansson-Stenman. 2021. "The Use of Green Nudges as an Environmental Policy Instrument." *Review of Environmental Economics and Policy* 15 (2): 216–237. [10]

Cerulli-Harms, Annette, Lorenz Goette, and Charles Sprenger. 2019. "Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences." *American Economic Journal: Microeconomics* 11 (1): 185–207. [38]

Chapman, Gretchen B., Helen Colby, Kimberly Convery, and Elliot J. Coups. 2015. "Goals and Social Comparisons Promote Walking Behavior." *Medical Decision Making* advance online publication. doi: 10.1177/0272989X15592156: [6]

Charness, Gary, and Uri Gneezy. 2009. "Incentives to Exercise." *Econometrica* 77 (3): 909–931. [40]

Chetty, Raj, John N. Friedman, Tore Olsen, and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics* 126 (2): 749–804. [37]

Clark, Damon, David Gill, Victoria Prowse, and Mark Rush. 2020. "Using Goals to Motivate College Students: Theory and Evidence From Field Experiments." *Review of Economics and Statistics* 102 (4): 648–663. [6, 10, 11]

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-González. 2015. "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Science* 61 (12): 2926–2944. [10]

Crawford, Vincent P., and Juanjuan Meng. 2011. "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income." *American Economic Review* 101 (5): 1912–1932. [11]

Creutzig, Felix, Joyashree Roy, William F. Lamb, Inês M. L. Azevedo, Wändi Bruine de Bruin, Holger Dalkmann, Oreane Y. Edelenbosch, Frank W. Geels, Arnulf Grubler, Cameron Hepburn, Edgar G. Hertwich, Radhika Khosla, Linus Mattauch, Jan C. Minx, Anjali Ramakrishnan, Narasimha D. Rao, Julia K. Steinberger, Massimo Tavoni, Diana Ürge-Vorsatz, and Elke U. Weber. 2018. "Towards demand-side solutions for mitigating climate change." *Nature Climate Change* 8 (4): 260–263. [9]

Della Vigna, Stefano. 2009. "Psychology and Economics : Evidence from the Field." *Journal of Economic Literature* 472: 315–372. [11]

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review* 96 (3): 694–719. [5]

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio. 2013. "Information Strategies and Energy Conservation Behavior: A Meta-analysis of Experimental Studies from 1975 to 2012." *Energy Policy* 61: 729–739. [10]

Diecidue, Enrico, and Jeroen van de Ven. 2006. "Aspiration Level, Probability of Success and Failure, and Expected Utility." *SSRN Electronic Journal*, [11]

Dietz, Thomas, Gerald T. Gardner, Jonathan Gilligan, Paul C. Stern, and Michael P. Vandenbergh. 2009. "Household Actions Can Provide a Behavioral Wedge to Rapidly Reduce US Carbon Emissions." *Proceedings of the National Academy of Sciences of the United States of America* 106 (44): 18452–18456. [9]

Dobronyi, Christopher R., Philip Oreopoulos, and Uros Petronijevic. 2019. "Goal Setting, Academic Reminders, and College Success: A Large-Scale Field Experiment." *Journal of Research on Educational Effectiveness* 12 (1): [10]

Dohmen, Thomas, Armin Falk, Klaus Fliessbach, Uwe Sunde, and Bernd Weber. 2011. "Relative versus absolute income, joy of winning, and gender: Brain imaging evidence." *Journal of Public Economics* 95 (3-4): 279–285. [11]

Drexler, Alejandro, Greg Fischer, and Antoinette Schoar. 2014. "Keeping It Simple: Financial Literacy and Rules of Thumb." *American Economic Journal: Applied Economics* 6 (2): 1–31. [5]

Drucker, Peter F. 1954. *The Practice of Management*. New York: Harper & Row. [5]

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

Ericson, Keith, and Andreas Fuster. 2011. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Quarterly Journal of Economics* forthcomin: [38]

Fan, James, and Joaquín Gómez-Miñambres. 2020. "Nonbinding Goals in Teams: A Real Effort Coordination Experiment." *Manufacturing & Service Operations Management* 22 (5): 1026–1044. [10]

Fang, Ximeng, Lorenz Goette, Bettina Rockenbach, Matthias Sutter, Verena Tiefenbeck, Samuel Schoeb, and Thorsten Staake. 2020. "Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Energy Conservation." *CRC TR 224 Discussion Paper No. 149*, [20, 26]

Farber, Henry S. 2005. "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers." *Journal of Political Economy* 113 (1): 46–82. [11]

Farber, Henry S. 2015. "Why you Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers*." *Quarterly Journal of Economics* 130 (4): 1975–2026. [11]

Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review* 97 (1): 298–317. [11, 42]

Ferraro, Paul J., and Michael K. Price. 2013a. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95 (1): 64–73. [10]

Ferraro, Paul J., and Michael K. Price. 2013b. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95 (1): 64–73. [22, 27]

Fischer, Corinna. 2008. "Feedback on Household Electricity Consumption: A Tool for Saving Energy?" *Energy Efficiency* 1 (1): 79–104. [10]

Frederiks, Elisha R., Karen Stenner, and Elizabeth V. Hobman. 2015. "Household Energy Use: Applying Behavioural Economics to Understand Consumer Decision-Making and Behaviour." *Renewable and Sustainable Energy Reviews* 41: 1385–1394. [5]

Gallus, Jana. 2017. "Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia." *Management Science* 63 (12): 3999–4015. [6]

Gerster, Andreas, Mark Andor, and Lorenz Goette. 2020. "Disaggregate Consumption Feedback and Energy Conservation." *CEPR Discussion Paper 14952*, [10]

Gneezy, Uri, Lorenz Goette, Charles Sprenger, and Florian Zimmermann. 2017. "The limits of expectations-based reference dependence." *Journal of the European Economic Association* 15 (4): 861–876. [38]

Goette, Lorenz, Thomas Graeber, Alexandre Kellogg, and Charles Sprenger. 2020. "Heterogeneity of Loss Aversion and Expectations-Based Reference Points." *Working Paper*, [38]

Goette, Lorenz, Hua-Jing Han, and Zhi Hao Lim. 2021. "The Dynamics of Goal Setting: Evidence From a Field Experiment on Resource Conservation." *CRC TR 224 Discussion Paper No. 283*, [26]

Gómez-Miñambres, Joaquín. 2012. "Motivation through goal setting." *Journal of Economic Psychology* 33 (6): 1223–1239. [9, 11]

Harding, Matthew, and Alice Hsiaw. 2014. "Goal setting and energy conservation." *Journal of Economic Behavior & Organization* 107: 209–227. [6, 10, 11]

Heath, Chip, Richard Larrick, and George Wu. 1999. "Goals as Reference Points." *Cognitive Psychology* 38: 79–107. [8, 9, 11, 38, 42]

Höpfner, Jessica, and Nina Keith. 2021. "Goal Missed, Self Hit: Goal-Setting, Goal-Failure, and Their Affective, Motivational, and Behavioral Consequences." *Frontiers in psychology* 12: 704790. [9]

Houde, Sebastien, Annika Todd, Anant Sudarshan, June A. Flora, and K. Carrie Armel. 2013. "Real-time Feedback and Electricity Consumption: A Field Experiment Assessing the Potential for Savings and Persistence." *The Energy Journal* 34 (1): [42]

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263–291. [11]

Karlin, Beth, Joanne F. Zinger, and Rebecca Ford. 2015. "The Effects of Feedback on Energy Conservation: A Meta-analysis." *Psychological Science* 141 (6): 1205–1227. [10]

Kleven, Henrik J., and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics* 128 (2): 669–723. [35]

Kleven, Henrik Jacobsen. 2016. "Bunching." *Annual Review of Economics* 8 (1): 435–464. [32, 35]

Koch, Alexander K., and Julia Nafziger. 2011. "Self-regulation through Goal Setting." *Scandinavian Journal of Economics* 113 (1): 212–227. [6, 9, 11]

Koch, Alexander K., and Julia Nafziger. 2016. "Goals and bracketing under mental accounting." *Journal of Economic Theory* 162: 305–351. [11]

Kollmuss, Anja, and Julian Agyeman. 2002. "Mind the Gap: Why Do People Act Environmentally and What Are the Barriers to Pro-Environmental Behavior?" *Environmental Education Research* 8 (3): 239–260. [5]

Kosfeld, Michael, and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3 (3): 86–99. [6]

Koszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121 (4): 1133–1165. [38, 40]

Koszegi, Botond, and Matthew Rabin. 2009. "Reference-Dependent Consumption Plans." *American Economic Review* 99 (3): 909–936. [38, 40]

Kuhn, Peter, and Lizi Yu. 2021. "Kinks as Goals: Accelerating Commissions and the Performance of Sales Teams." *IZA Discussion Paper No. 14115*, [11]

Locke, Edwin A., and Gary P. Latham. 1990. *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice-Hall. [5, 8, 10, 11, 26]

Locke, Edwin A., and Gary P. Latham. 2002. "Building a practically useful theory of goal setting and task motivation. A 35-year odyssey." *The American psychologist* 57 (9): 705–717. [5, 10, 11]

Locke, Edwin A., and Gary P. Latham. 2013. *New developments in goal setting and task performance*. New York: Routledge. [11]

Locke, Edwin A., and Gary P. Latham. 2019a. "Does prospect theory add or subtract from our understanding of goal directed motivation?" In *The Only Constant in HRM Today is Change*. Edited by Diana L. Stone and James H. Dublebohn. Charlotte, NC: Information Age Publishing, 19–42. [11]

Locke, Edwin A., and Gary P. Latham. 2019b. "The development of goal setting theory: A half century retrospective." *Motivation Science* 5 (2): 93–105. [10]

Loock, Claire-Michelle, Thorsten Staake, and Frédéric Thiesse. 2013. "Motivating Energy-Efficient Behavior With Green IS: An Investigation of Goal Setting and the Role of Defaults." *MIS Quarterly* 37 (4): 1313–1332. [6]

Madrian, Brigitte C. 2014. "Applying Insights From Behavioral Economics To Policy Design." *Annual Review of Economics* 6: 663–688. [9]

Markle, Alex, George Wu, Rebecca White, and Aaron Sackett. 2018. "Goals as reference points in marathon running: A novel test of reference dependence." *Journal of Risk and Uncertainty* 56 (1): 19–50. [11]

Mento, Anthony J., Robert P. Steel, and Karren, Ronald, J. 1987. "A Meta-Analytic Study of the Effects of Goal-Setting on Task Performance: 1996-1984." *Organizational Behavior and Human Decision Processes* 39: 52–83. [10]

PUB. 2018a. "Singapore World Water Day 2018 & Household Water Consumption Study." *Press Release by Singapore's National Water Agency* March 1: [6]

PUB. 2018b. "Smart Shower Programme." *https://www.pub.gov.sg/savewater/athome/smartshowerprogramme (accessed June 7, 2018)*, [43]

Reddy, Sheila M.W., Jensen Montambault, Yuta J. Masuda, Elizabeth Keenan, William Butler, Jonathan R.B. Fisher, Stanley T. Asah, and Ayelet Gneezy. 2017. "Advancing Conservation by Understanding and Influencing Human Behavior." *Conservation Letters* 10 (2): 248–256. [9]

Samek, Anya. 2019. "Gifts and goals: Behavioral nudges to improve child food choice at school." *Journal of Economic Behavior & Organization* 164: 1–12. [10]

Schmitt, Kathrin, Verena Tiefenbeck, Ximeng Fang, Lorenz Goette, Thorsten Staake, and Davin Wang. 2021. "Pro-environmental spillover effects in the resource conservation domain: Evidence from a randomized controlled trial in Singapore." *mimeo*, [43]

Sydnor, Justin. 2010. "(Over)insuring Modest Risks." *American Economic Journal: Applied Economics* 2 (4): 177–199. [42]

Taylor, Michael, and Claudio Accheri. 2019. "This is Singapore's plan to avoid running out of water." *World Economic Forum* August 13 (https://www.weforum.org/agenda/2019/08/singapore-focus-innovation-key-securing-water-future): [6]

Thakral, Neil, and Linh T. Tô. 2021. "Daily Labor Supply and Adaptive Reference Points." *American Economic Review* 111 (8): 2417–2443. [11, 40]

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth and Happiness*. Yale University Press. [9]

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake. 2018. "Overcoming salience bias: how real-time feedback

fosters resource conservation." *Management Science* 64 (3): 1458–1476. [6, 7, 10, 14, 20, 22, 26–28]

Tiefenbeck, Verena, Anselma Woerner, Samuel Schoeb, Elgar Fleisch, and Thorsten Staake. 2019. "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives." *Nature Energy* 4: 35–41. [10]

Wood, Wendy, and Dennis Rünger. 2016. "Psychology of Habit." *Annual review of psychology* 67: 289–314. [40]

Chapter 2

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

Joint with Lorenz Goette, Bettina Rockenbach, Matthias Sutter, Verena Tiefenbeck, Samuel Schoeb, and Thorsten Staake

2.1 Introduction

Amidst growing public 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, Stenner, and Hobman, 2015). This gap between intentions and actions can result from a multiplicity of behavioral frictions and biases. For instance, previous research has shown that individuals tend to underestimate the impact of highly resource-intensive behaviors (Attari, DeKay, Davidson, et al., 2010; Attari, 2014), and that they may also not be fully attentive to their resource use (Allcott, 2016; Tiefenbeck, Goette, Degen, et al., 2018).

Other factors such as self-control problems and status quo bias may certainly also play a role. Importantly, however, such behavioral biases could not only prevent consumers from acting on their intrinsic prosocial or pro-environmental motivations, but also mute their response to policy interventions aimed at encouraging behavioral change. Thus, when multiple dimensions of bias are present at the same time, interventions that miss an important dimension may fail to unfold their full potential. For example, providing information about environmental impacts may have little effect on behavior if individuals remain inattentive to their resource use.[1] Conversely,

1. Information provision 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,

making resource use salient may only have a muted effect if agents remain unaware of adverse environmental impacts. Hence, in this example, a combined approach that targets both imperfect information and inattention could have synergetic, mutually reinforcing effects, i.e. positive interaction effects or complementarities. More generally, we argue that bundling interventions can result in complementarities if each intervention is particularly suited to address a different source of behavioral bias. Following Coe and Snower (1997), we define interventions as complements if *each* intervention becomes more effective when implemented in conjunction with the other(s) than in isolation. While many studies consider the use of combined interventions, there is need for more theoretical and empirical research that investigates systematic drivers of complementarity (or substitutability) and thereby provides guidance for the design of effective behavioral policy.

In this paper, 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 we target different potential sources of behavioral bias against resource conservation. 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 or CO2 emissions — through a smart meter display (Tiefenbeck, Goette, 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, which allows us to uncover potential complementarities.

To formalize our argument as to why complementarities might arise in such a context, we introduce a stylized theoretical framework in which biased perceptions of resource use arise from multiple sources (e.g. imperfect information, limited attention). Each of these biases acts akin to a discount factor and thus prevents agents from fully incorporating the marginal costs of resource consumption into their behavior. A key prediction from our framework is that when each bias mutes the perceived cost of resource use independently of other biases, then the effects of pro-environmental interventions that mitigate different sets of biases can reinforce each other, so that the interventions become complements. The intuition is simple: the more one particular bias is reduced, the larger is the impact of reducing another bias. 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

Larrick, Hossain, et al., 2019) and tend to engage in relatively ineffective conservation measures (Gardner and Stern, 2008; Tonke, 2019).

the environmental impact is more negative than previously thought. This interaction mechanism is absent when two interventions mostly operate through the same behavioral channel, e.g. if they both provide the same type of information.

There are several reasons why (warm) water consumption in the shower provides an interesting context for studying complementarities in behavioral interventions. First, showering is a resource-intensive activity: an average shower in our sample requires 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.[2] Second, individuals tend to underestimate the CO_2 emissions caused by warm water consumption in the shower — by as much as 89% on average based on one of our surveys —, which creates scope for reducing energy consumption through information provision (Byrne, La Nauze, and Martin, 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 (Tiefenbeck, Goette, et al., 2018). Since individuals may not 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, it may be necessary to draw on both of these mechanisms at the same time.

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, with 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 (installed directly below the shower head) that recorded detailed data of each water extraction. 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 a baseline stage of 10 showers, the smart meter started displaying real-time feedback on water use for subjects in RTF and DUAL. About halfway into the study, we further started constructing the individualized energy reports using uploaded data from the smart meters and sent them out to subjects in SER and DUAL via email. This staggered design allows us to identify and estimate treatment effects of each intervention regime in a difference-in-differences setup. The shower energy reports mainly aimed at reducing knowledge gaps about environmental impacts, whereas real-time feedback mainly aimed at focusing attention and creating a sense

2. Calculated based on information from the German Federal Statistical Office. Source: https://www.destatis.de/EN/Themes/Society-Environment/Environment/_node.html

of immediacy. As both mechanisms might be important for encouraging conservation behavior, we hypothesize that the two interventions are complements.[3]

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 effects. 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 0.22 kWh of energy (3.8 liters of water) per shower, i.e. by more than 50%. Hence, the shower energy reports simply appeared to require an enhanced choice environment to become effective. The additional reduction of energy 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. Furthermore, we generally find no evidence of adjustments on the extensive margin, i.e. the number of showers people take. One noteworthy feature of our sample is that subjects had no monetary incentives for conserving energy or water, since they paid a flat fee for utilities. Thus, all conservation effects are driven solely by non-monetary motives, which makes them even more remarkable.

Additional questionnaire data shows that both interventions helped subjects form more precise beliefs about their own water use in the shower; 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 is not 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 hypothesis that the presence of multiple bias dimensions can induce complementarities between interventions that largely operate through different behavioral mechanisms. This implies that appropriate policy bundling may increase the cost-effectiveness of interventions beyond what can be achieved with piecemeal approaches. In particular, lack of evidence for

3. 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 sources. For example, real-time feedback could be interpreted as information provision about instantaneous water consumption that can facilitate learning or optimization, and this information can be complementary to the information on CO_2 emissions provided through shower energy reports.

effectiveness of an intervention in isolation — as for information provision through shower energy report in our case — does not imply that it cannot be effective in an enhanced policy environment that also takes into consideration further potential sources of bias.

Our study builds on important previous contributions that have studied the effects of similar behavioral interventions on household energy 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 a household-level conservation effect of 2%, or about 0.62 kWh per day; effectivity might be smaller outside the U.S., where the baseline energy consumption tends to be lower (see e.g. Andor, Gerster, Peters, et al., 2020, for a sample of German households), or when there are little monetary incentives to save energy (Myers and Souza, 2019). Our SER intervention also gives feedback about past consumption patterns, although differing to classical home energy reports in several aspects, mainly in that it targets one specific activity (showering) instead of aggregate household consumption. Disaggregated, activity-specific feedback could enable better learning and thus stronger conservation responses in the targeted activities (Gerster, Andor, and Goette, 2020), in particular when provided in shorter time intervals or even in real time. Tiefenbeck, Goette, 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 22% conservation effect, or, in absolute terms, a reduction of 0.6 kWh energy and 9 liters of water per shower. These results also replicate in a sample without monetary incentives and without self-selection into the study (Tiefenbeck, Woerner, Schoeb, et al., 2019). As real-time feedback can make resource consumption immediately salient, a natural question is whether we can use this to improve the effectiveness of other interventions that aim to encourage conservation behavior through further mechanisms like more detailed information provision or social norms and could thus benefit from generally higher attention to pro-environmental motives.

We further relate to a number of other studies that test a combination of different interventions, 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.[5] For example, Jessoe and Rapson (2014) find that pricing schemes that incentivize lower peak electricity consumption can fail to change be-

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, Steg, Vlek, et al. (2005), Fischer (2008), Delmas, Fischlein, and Asensio (2013), Karlin, Zinger, and Ford (2015), Andor and Fels (2018), Carlsson, Gravert, Kurz, et al. (2021).

5. 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, Karlan, and Zinman, 2014), education (Mbiti, Muralidharan, Romero, et al., 2019), risky sexual behavior (Duflo, Dupas, and

havior due to consumers not knowing how to effectively adjust electricity usage; only households who have been outfitted with in-home-displays reduce consumption significantly in response to price hikes. 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, Metcalfe, Price, et al., 2017; Holladay, LaRiviere, Novgorodsky, et al., 2019; Giaccherini, Herberich, Jimenez-Gomez, et al., 2020; Fanghella, Ploner, and Tavoni, 2021). Hahn, Metcalfe, Novgorodsky, 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, List, Metcalfe, 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. While we add to this literature by providing a novel case study on the complementarity of two specific types of behavioral interventions, our main contribution is that we attempt to make a step towards understanding mechanisms that systematically lead different policy interventions to become complements or substitutes. Hence, the empirical design is embedded within a conceptual framework — highlighting specifically the role of multiple sources of behavioral bias — that can be adapted to form hypotheses about policy interactions in other contexts as well.

The remainder of this paper is structured as follows: Section 2.2 introduces the theoretical framework for policy interactions under multiple sources of behavioral bias. Section 2.3 describes the experimental setup and derives behavioral predictions. Section 2.4 presents our data as well as some descriptive statistics. Section 2.5 explains our empirical approach and Section 3.3 presents our main empirical results. In Section 2.7, we study the potential mechanisms underlying the results and provide robustness checks. Section 3.4 concludes.

Kremer, 2015; Dupas, Huillery, and Seban, 2018), demand for health products (Ashraf, Jack, and Kamenica, 2013), or immunization (Banerjee, Chandrasekhar, Dalpath, et al., 2021). Many of these studies, however, cannot explicitly test policy interactions, 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, Chandrasekhar, 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.

2.2 Theoretical framework

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

2.2.1 Setup

Basic setup. — The agent engages in an energy-intensive activity, say showering and the policy objective is to reduce energy use. Her consumption level is determined by a trade-off between the consumption utility (incl. pleasure, instrumental benefits, opportunity costs of time) and the perceived costs of resource use (incl. monetary costs, environmental concern). She chooses energy use level $e \geq 0$ to maximize

$$U(e) \;=\; V(e) \,-\, B \cdot C(e)\,, \tag{2.2.1}$$

where $V(e)$ is the instantaneous consumption utility and $C(e)$ is the cost of energy consumption.[6] In addition to standard smoothness conditions, we assume that V is hump-shaped (locally increasing at 0, strictly concave, unique maximum) and that C is strictly monotonically increasing and weakly convex. For simplicity, we abstract from uncertainty or dynamics. In the absence of monetary motives, as in our empirical setting, $C(e)$ is the "moral" cost the agent perceives in face of the negative externalities from energy use. However, the cost function is attenuated by an aggregate bias factor B, and energy use is biased upwards if $B \in [0, 1)$.

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

$$B \;=\; b_1 \cdot b_2\,. \tag{2.2.2}$$

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, Goette, et al., 2018). The multiplicative form captures that any single factor can independently prevent the agent from implementing her conservation motive. 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

6. The agent may not explicitly optimize with regard to energy use, but as long as the mapping from actual decision variable (e.g. shower duration) to resource use is injective, we can represent the problem *as if* the agent was optimizing over energy use.

itself is sufficient.[7] Note that the entire framework can be easily generalized to the case of $B = \prod_{k=1}^{K} b_k$ with $k > 2$.

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

$$V'(e) \quad = \quad B \cdot C'(e)\,. \tag{2.2.3}$$

If $B < 1$, then the marginal cost is underweighted and energy use is thus biased upwards. Defining f such that $f(e) = \frac{V'(e)}{C'(e)}$ for all $e \in [0, \infty)$, we can directly map the relation between implemented energy use and aggregate bias as

$$e(B) \quad = \quad f^{-1}(B)\,, \tag{2.2.4}$$

because equation (2.2.3) implies that $f(e(B)) = B$. Notice that f^{-1} is a strictly decreasing function, so the weaker the aggregate bias, i.e. B closer to 1, the lower the energy use.[8] In this sense, B can be interpreted as an input for energy conservation.

Behavioral interventions. — In this setup, we define behavioral interventions as policies that aim to change consumers' behavior by changing B.[9] In contrast, price-based policies would be aimed at increasing the marginal costs of energy use, $C'(e)$, that the agent faces.[10] 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).

2.2.2 Policy interaction mechanisms

Two interventions, X and Y, are complements if their combination reduces bias by more than the sum of their individual effects, i.e $\Delta B^{X+Y} > \Delta B^X + \Delta B^Y$. If they are

7. This is reminiscent of the Anna Karenina principle, which states that failure in a single factor may lead to failure of an endeavor as a whole. It is inspired by the opening phrase of Leo Tolstoy's novel *Anna Karenina*: "Happy families are all alike; every unhappy family is unhappy in its own way." (Tolstoy, 2003).

8. This is because marginal consumption utility $V'(e)$ is strictly decreasing and marginal cost $C'(e)$ is non-decreasing. Hence, f is strictly increasing, so the inverse function f^{-1} exists and is strictly decreasing.

9. Equation (2.2.4) shows that any policy X that mitigates the aggregate bias (B^X) compared to no-intervention state B will induce the agent to conserve energy. Hence, $\Delta B^X = B^X - B > 0$ implies that $\Delta e^X = e(B^X) - e(B) < 0$. The more successful an intervention is in mitigating the aggregate bias, the larger the energy reduction effect.

10. Our framework also allows for an interpretation that takes more a social planner's point of view, aiming for the agent to internalize the full social cost $C^s(e)$. The ratio of private to social cost $C(e)/C^s(e)$ would then be another factor entering into the aggregate bias B^s, so decision utility is $U(e) = V(e) - B^s \cdot C^s(e)$. This interpretation highlights the overarching policy objective of reducing externalities instead of "internalities". Efforts to increase the privately perceived cost can include Pigouvian taxes (e.g. carbon pricing), social norms, goal-setting, etc.

substitutes, the inequality sign is reversed. Notice that even under substitutability, it can be the case that $X + Y$ is more effective than either X or Y in isolation, i.e. $\Delta B^{X+Y} > \Delta B^{X}$ and $\Delta B^{X+Y} > \Delta B^{Y}$. Thus, to empirically identify interaction effects between different policy interventions, it is also necessary to evaluate the effectiveness of each intervention in isolation. Our theoretical framework allows for several mechanisms that could make interventions either complements or substitutes.

Complementary policy levers. — The key mechanism we aim to highlight in this paper is that in the presence of multiple sources of bias, policies that target only one dimension may have a limited effect on behavior, whereas the effect of combining several policy levers may be superadditive. This is an immediate consequence of the multiplicative structure of B, which implies a positive cross-derivative: $(\partial^2 B / \partial b_1 \partial b_2) > 0$. 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 (2.2.2), can be thought of as the area of a rectangle with sides of lengths b_1 and b_2 (see Figure 2.2.1a). 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

$$\Delta B \quad = \quad \delta_1 b_2 \; + \; \delta_2 b_1 \; + \; \delta_1 \delta_2 \, . \tag{2.2.5}$$

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 right rectangle outlined in the second graph of Figure 2.2.1. This mechanism potentially induces complementarity between interventions that are specialized on mitigating different sources of bias each.

In practice, it may be hard to design "pure" interventions where each intervention changes only one dimension of B. 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, Mullainathan, and Schwartzstein, 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 informa-

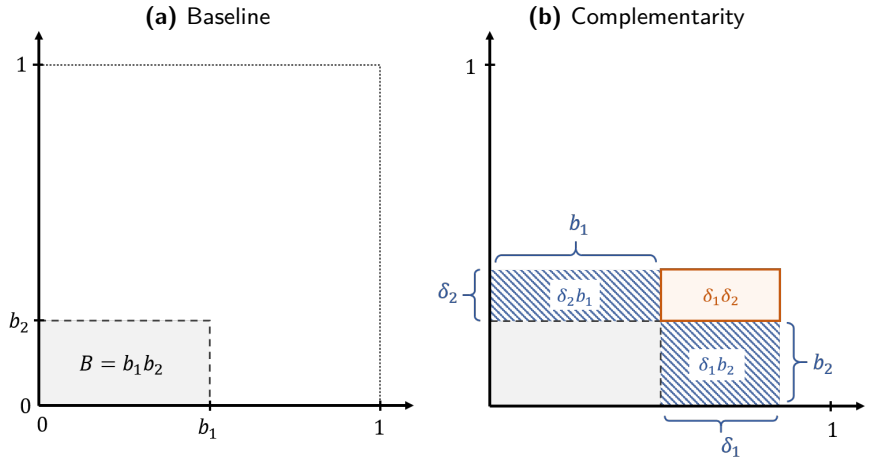


Figure 2.2.1. Depiction of example interventions

Notes: Notes. The grey rectangle in Figure (a) illustrates the aggregate bias B as defined in equation 2.2.2 without any intervention in place. Figure (b) illustrates the increase in B through exogenous interventions in each dimension.

tion 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 2.2.2a 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 2.2.2c 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

$$\Delta B^{X+Y} - \Delta B^X - \Delta B^Y = \delta_1^X \delta_2^Y + \delta_2^X \delta_1^Y. \tag{2.2.6}$$

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. $\delta_i^{X+Y} = \max(\delta_i^X, \delta_i^Y)$. This is illustrated in Figure 2.2.2d. This case is less favorable toward complementarities, as each intervention now only has an impact on one bias dimension, and the condition becomes

$$\Delta B^{XY} - \Delta B^X - \Delta B^Y = (\delta_1^X - \delta_1^Y)(\delta_2^Y - \delta_2^X) - (\delta_2^X b_1 + \delta_1^Y b_2 + \delta_2^X \delta_1^Y) \tag{2.2.7}$$

This term is positive if the top right rectangle in Figure 2.2.2d, 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

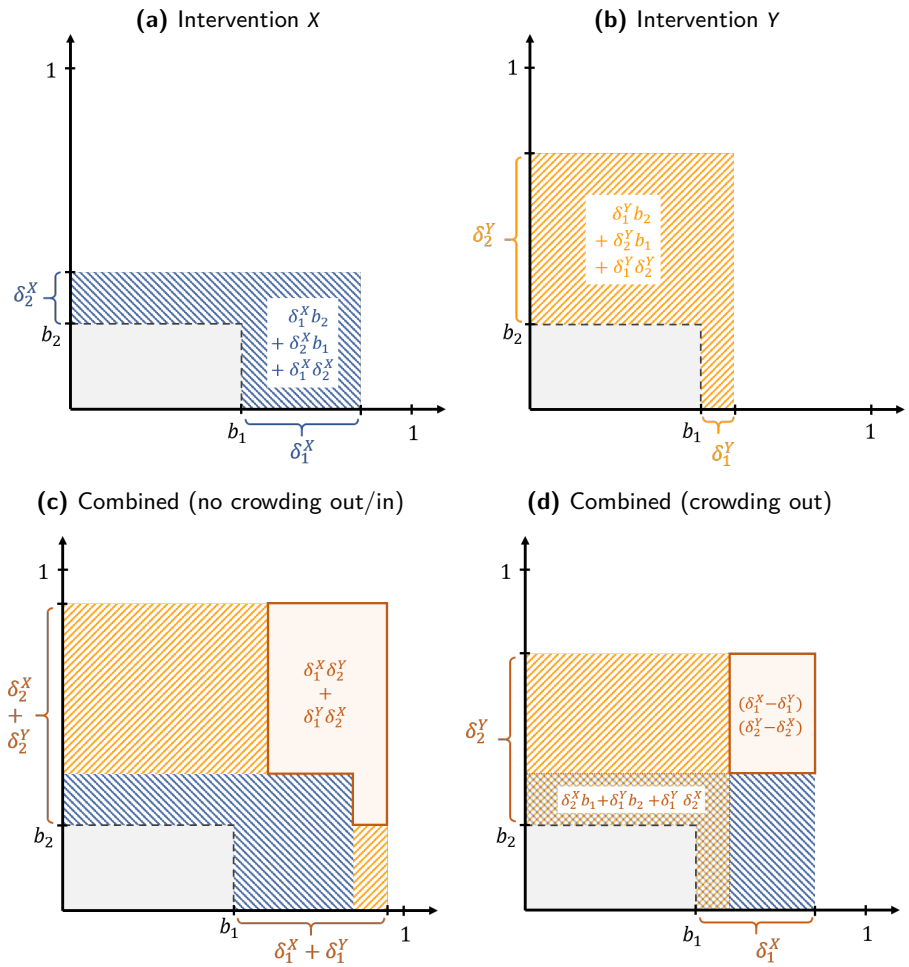


Figure 2.2.2. Depiction of example interventions

Notes: 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.

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.

Complementarity in behavioral outcomes. — So far we have focused on mechanisms of complementarity in manipulating B. How this maps into observable behavior depends on the mapping of B to e. The condition for overall policy complementarity in the outcome of interest, energy consumption, can be written as

$$\Delta e^{X+Y} \leq \Delta e^X + \Delta e^Y \tag{2.2.8}$$

Typically, one would expect a decreasing responsiveness, as resource consumption is more inelastic at lower levels (e.g. due to a desire for satisfying basic needs like

hygiene), so the scope for further conservation effects diminishes with every intervention that is piled upon another. In our framework, this corresponds to function f^{-1} being convex.[11] Intuitively, the more the agent already reduces her consumption, the more difficult it becomes to further reduce it. Thus, under this assumption, observing complementarities in behavioral outcomes implies complementarities in bias mitigation.

Empirically identifying complementarities is important for optimal policy. Consider the following stylized application: A policy maker has the objective of reducing the average energy consumption $\bar{e}$ in the population and has at her disposal two equally-costly and equally-effective behavioral interventions X and Y. Suppose that the budget allows for treating fraction $\alpha \in (0,1)$ of the population with an intervention. Alternatively, the policy maker can also treat $\alpha/2$ of the population with a combined intervention $X+Y$. The latter is (weakly) superior to the former precisely when the complementarity condition in equation (2.2.8) holds. Thus, it is important to study empirically whether two interventions are complements or substitutes.

2.3 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.

2.3.1 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. These students have no direct monetary incentives to conserve energy or water, because they pay a flat monthly

11. For example, if V has a positive third derivative and the cost function C is linear or quadratic, then f^{-1} is strictly decreasing and convex. A positive third derivative is often labeled prudence and implies a desire for precautionary saving in choice under risk. Of course, f^{-1} could in principle also be concave, so marginal returns are increasing, but this seems implausible. For example, concavity can imply that conservation programs have larger effects for low-baseline consumers, although the opposite is usually true.

rent that includes all utility bills. Hence, any observed conservation response would be solely driven by non-monetary motives and unconfounded by income effects.

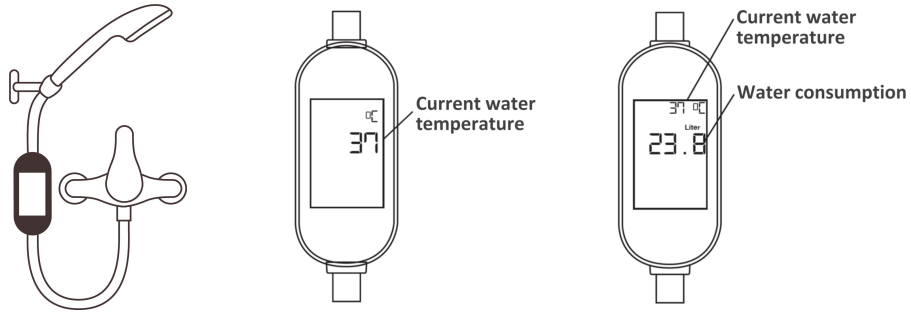
To participate in the study, residents had to actively agree based on the principle of informed consent. Two additional criteria were levied: subject should not have lengthy absences planned within the intended study period (except during Christmas vacation), and they should own a smartphone compatible 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.[12] 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.

2.3.2 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 2.3.3a). 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. After water flow is stopped, the device remains powered on for three minutes with the display remaining active. If the water is turned on again within this time frame, the device will continue measurement from the point where it had previously stopped. This accounts for short breaks

12. 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 2.3.3. Amphiro b1 smart shower meter

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 data as the most recent observation 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.[13] 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 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.

2.3.3 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

13. 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.

study (Figure 2.3.3b). 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 2.3.3c).

2.3.4 Implementation of shower energy reports

Our second type of 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 2.A.1 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 whole 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 the original Opower home energy reports, see Appendix Figure 2.A.2. Specifically, we assigned a random anonymous peer to each subject and displayed statistics on the peer's energy and water use.[14] At the bottom of each report, there was a personalized link to a mini-survey that we can use to verify if, and how closely, the email has been read.

2.3.5 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

14. 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.

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.[15]

Figure 2.3.4 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 behavior. 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 permanently 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 23 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 the first report was sent out.[16] 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 were sent out. These subjects were simply asked to fill out a mini-survey, the same 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 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 the "true" counterfactual, as subjects still receive a smart meter with temperature information and placebo emails instead of shower energy report. 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.

15. For the exact randomization protocol, see Appendix B.

16. 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.

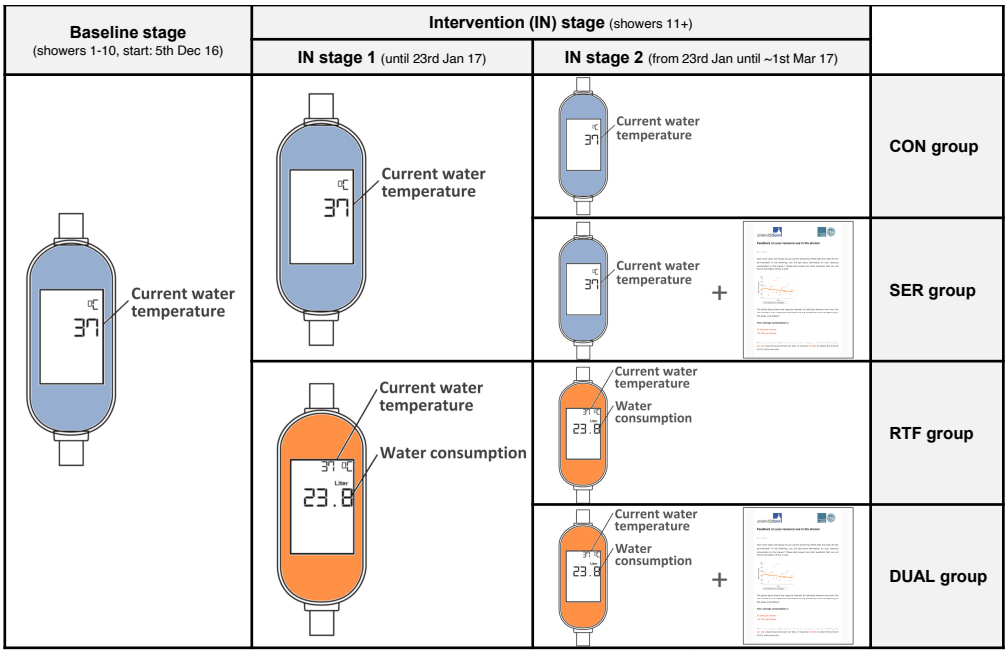


Figure 2.3.4. Experimental design and timing of interventions

2.3.6 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, Goette, 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 comparison with a randomly assigned and anonymous peer, which might further add motivation through social norms, although Tiefenbeck, Goette, et al. (2018) find no effect of including comparisons with the co-resident in a two-person household. However, the reports are not immediately salient while showering.

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

The conservation effect of knowledge gains through energy reports could be stifled by remaining barriers like limited attention or self-control problems, which can be more suitably targeted by real-time feedback.[17] 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 the theoretical framework, 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.

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 (treatment DUAL) than for subjects who do not receive real-time feedback (treatment SER).*

2.4 Data and descriptive statistics

2.4.1 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

17. 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).

(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.[18] 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.[19] 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, Goette, 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 produce 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.[20] 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 bounds for when a shower

18. 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, Goette, et al. (2018), we also assume boiler efficiency losses of 35% and distribution losses of 24%.

19. 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.

20. 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.

took place. We instructed subjects to use the smartphone app regularly starting from 11 January 2017, and sent additional reminders before each energy report was sent out. Using this timing information, we classify observations into pre-report showers (IN stage 1) or post-report showers (IN stage 2). If there are multiple showers within the range of uncertainty around report dates, we use the switching point implied by constant shower frequency. One complication is that we do not know the timing of showers by the subjects who did not manage to upload any data to the app. Therefore, we impute the timing of showers for these non-uploaders based on the assumption that timing of shower energy reports 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, and then assign the implied post-report probabilities to showers of non-uploaders. Figure 2.A.3 in Appendix A plots the estimated CDFs.[21]

2.4.2 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”.[22] Shower comfort is elicited using five items on how much subjects enjoy showering, e.g. “I find it relaxing to take a shower”.[23] 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 di-

21. For more details of the imputation procedure, see Appendix D.

22. 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, Schultz, Cialdini, et al. (2008) in the baseline questionnaire.

23. 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 2.4.1. 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	2489
Volume [liter]	37.82	30.45	9.20	29.60	76.00	2489
Duration [min]	7.00	5.01	1.96	5.83	13.01	2489
Temperature [Celsius]	36.16	5.22	32.00	37.00	40.00	2463
Flow rate [l/min]	5.71	2.45	2.80	5.40	9.10	2489

Includes only showers taken in the baseline stage, i.e. first 10 showers and before shower energy reports were sent out. 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.

rectly compared to their actual water use as measured by the smart meter. Note that we refrained from eliciting subjects' beliefs about 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.

2.4.3 Sample characteristics and baseline behavior

All participants in the field experiment 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 share of females was 61 percent.[24] 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 2.4.1 presents descriptive statistics about baseline energy and water use per shower, as well as shower duration (net of breaks), water temperature, and flow rate. Shower duration is calculated from dividing water volume by average flow rate. On average, showers in the baseline stage feature 7 minutes of water flow, which amounts to 37.82 liters of water. On average, water is heated up to a temperature of 36.16 degrees Celsius, resulting in energy use 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. The average flow rate of 5.74 liters per minute is relatively low, likely due to dorm infrastructure not being

24. In 2016/17, the overall share of female students was 55% at the University of Bonn and 60% at the University of Cologne, suggesting that there was no substantial gender-based selection into our study.

Table 2.4.2. 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
SER group	-0.066	-1.901	0.181	0.959	-0.435	3.393
	(0.220)	(3.468)	(0.548)	(0.608)	(0.320)	(5.226)
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)
DUAL group	-0.057	-0.910	0.213	0.320	-0.165	3.224
	(0.226)	(3.575)	(0.581)	(0.560)	(0.358)	(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	0.966	0.958	0.969	0.356	0.571	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.

up to modern standards (flow rates of 10-12 liters per minute are more typical for German households).

2.4.4 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 2.4.2 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.

2.4.5 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 2.4.2, 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.

2.4.6 Presence of imperfect information and behavioral biases

Before moving on to the analysis of our experimental interventions, we provide some descriptive evidence that individuals' resource consumption in our setting may indeed be subject to significant behavioral frictions 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 2.4.1 shows that subjects' estimates are all over the place, and we cannot reject the null hypothesis that estimated and measured water use are in fact uncorrelated (Pearson's $\rho = 0.08, p = 0.1825$). This clearly demonstrates that subjects were not well informed about their own behavioral outcomes prior to any intervention.[25] Interestingly, however, the mean estimate across all subjects (39.8 liters) is close to the actual mean water use per shower in the baseline stage (37.8 liters). 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.

Furthermore, subjects are probably especially unaware of how much energy is consumed, and hence CO_2 emitted, in a typical shower. 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 ($n = 329$). 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

25. 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 extreme outliers with estimates above 200 liters. The corresponding regression results are presented in Appendix A Table 2.A.5.

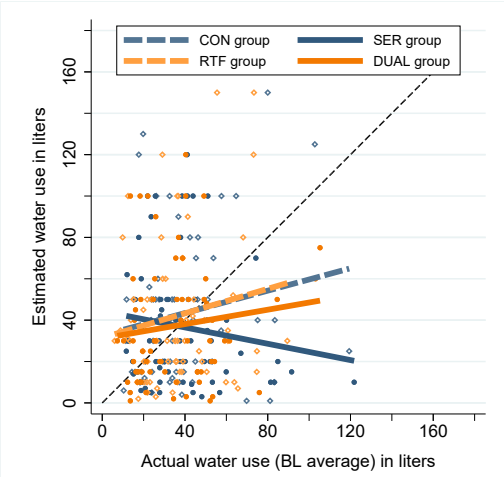


Figure 2.4.1. Pre-intervention awareness about water use per shower

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.

for the amount of water used per shower was fairly unbiased.[26] Thus, there might be a large potential for encouraging energy conservation through the information provided in shower energy reports (Byrne, La Nauze, and Martin, 2018).

Although anecdotally compelling, finding direct evidence for inattention or self-control problems in the shower is trickier. The closest 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." on a five-point Likert scale. 59% of our sample states that they agree or strongly agree to the statement (34% agree, 25% strongly agree), whereas only 18% of subjects disagree or strongly disagree (13% disagree, 5% strongly disagree), indicating that a lack of focus while showering is prevalent. We find that subjects' response to this item is significantly correlated with their baseline energy use in the shower (Pearson's $\rho = 0.17, p = 0.003$). In fact, it is the single most predictive item for baseline consumption in the entire survey. Our interventions could thus help reduce energy use by reminding subjects not to lose track of time completely under the shower.

26. On average, students estimated that a typical shower causes emissions of 91.3 grams of CO_2 (median 35 grams). The actual emissions amount based on the data from our experiment is about 800 grams. 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. For more details, see Appendix E.

2.5 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.

2.5.1 Basic estimation strategy

To formally estimate the effects of different intervention regimes, we exploit the staggered introduction of real-time feedback and shower energy reports in the experimental design, which gives us a double-layered difference-in-differences setup. The differential changes in consumption behavior 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).

For estimating the effect of real-time feedback in isolation, the most straightforward and easy-to-interpret approach is to simply compare subjects in the RTF and CON groups over the entire experimental period, as these subjects never received shower energy reports in any form. We do so by estimating the equation

$$y_{it} \;=\; \alpha_i + \beta_0 IN_{it} + \beta_1 IN_{it} \times T_i^R + \varepsilon_{it}\,, \tag{2.5.1}$$

where the outcome variable y_{it} is energy use (water use) by individual i for shower number t, α_i is the individual fixed effect, IN_{it} is an indicator that takes the value 1 if observation it falls into the intervention stage (i.e. $t > 10$), and T_i^R is an indicator for being assigned to treatment group RTF. The coefficient of interest is β_1, which corresponds to the average treatment effect of real-time feedback (in isolation) over the entire three months of the study. In this specification, we do not have to deal with issues relating to non-compliance and timing of reports, though it comes at the cost of disregarding half of the sample in intervention stage 1.

To make use of the full sample when estimating the effect of real-time feedback, we can compare differential changes in consumption behavior from baseline stage to intervention stage 1 for the pooled RTF/DUAL group versus the pooled CON/SER group, because real-time feedback had already phased in but shower energy reports had not. For intervention stage 2, when shower energy reports started flying in, we split up the pooled groups again, so the regression equation is

$$\begin{aligned} y_{it} \;=\; \alpha_i \;&+\; IN_{it} \times \left(\beta_0 \;+\; \beta_1 T_i^{R/D}\right) \\ &+\; IN_{it}^{s2} \times \left(\gamma_0 \;+\; \gamma_1 T_i^{R/D} \;+\; \gamma_2 T_i^S \;+\; \gamma_3 T_i^D\right) \;+\; \varepsilon_{it}\,. \end{aligned} \tag{2.5.2}$$

IN_{it} is again the indicator for the intervention stage, and IN_{it}^{s2} is an indicator for showers that fall into intervention stage 2 (post-report). $T_i^{R/D}$, T_i^D and T_i^S are treatment group indicators, where superscript R/D denotes the combined groups RTF and DUAL, superscript D denotes the DUAL group, and superscript S denotes the SER group only. As IN_{it} remains switched on for the entire intervention period, IN_{it}^{s2} comes on top of that, so all γ-parameters need to be interpreted as incremental changes from intervention stage 1 to intervention stage 2.

Equation (2.5.2) incidentally also includes estimates for the effect of shower energy reports (γ_2 and γ_3), but one concern here is that they do not control for differences between RTF and DUAL or between CON and SER in the first intervention stage. Although the pooled groups in intervention stage 1 should behave the same before reports are sent out, some random differences are likely to exist, and these would propagate to the estimates of γ_2 and γ_3. For estimating the effects of shower energy reports we therefore prefer the more flexible model in which treatment groups are considered separately from the beginning of the intervention stage:

$$\begin{aligned} y_{it} \quad = \quad \alpha_i \quad + \quad IN_{it} \times \left(\beta_0 \; + \; \beta_1 T_i^{R/D} \; + \; \beta_2 T_i^S \; + \; \beta_3 T_i^D\right) \\ + \quad IN_{it}^{s2} \times \left(\gamma_0 \; + \; \gamma_1 T_i^{R/D} \; + \; \gamma_2 T_i^S \; + \; \gamma_3 T_i^D\right) \quad + \quad \varepsilon_{it}\,. \end{aligned} \tag{2.5.3}$$

Given the model formulation, 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, while γ_1 is the change in treatment effect in the second stage. γ_2 is the treatment effect of shower energy reports in isolation, and γ_3 is the additional effect of adding shower energy reports to real-time feedback. The relevant comparisons of interests are between SER and CON on the one hand — for the effect of reports without real-time feedback — and between DUAL and RTF on the other hand — for the marginal effect of adding reports to reinforce the already existing real-time feedback.

2.5.2 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).[27] 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

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

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.

One possible approach to estimate treatment effects under imperfect treatment take-up is to run an intention-to-treat (ITT) analysis, which ignores that some participants did not actually receive informative shower energy reports and simply uses treatment assignment to estimate treatment effects. However, this is not very appealing in our context, as failure of information provision due to technical problems is in principle an avoidable problem. The more policy-relevant treatment effect is the effect of delivering informative shower energy reports. Therefore, our preferred approach is to estimate the treatment effect on the treated (TOT), i.e., on subjects who managed to upload data and thus received actual information through the shower energy reports.

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.

The second way in which we estimate the TOT is by using random treatment assignment as instrument for actual take-up.[28] 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).[29] Compared to the "uploaders-only"-approach, the instrumental variables approach is always consistent, but potentially inefficient. We will report the results from both TOT-approaches, but the estimates are very similar, suggesting that endogeneity is not a large issue in our setting.

28. 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.

29. 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.

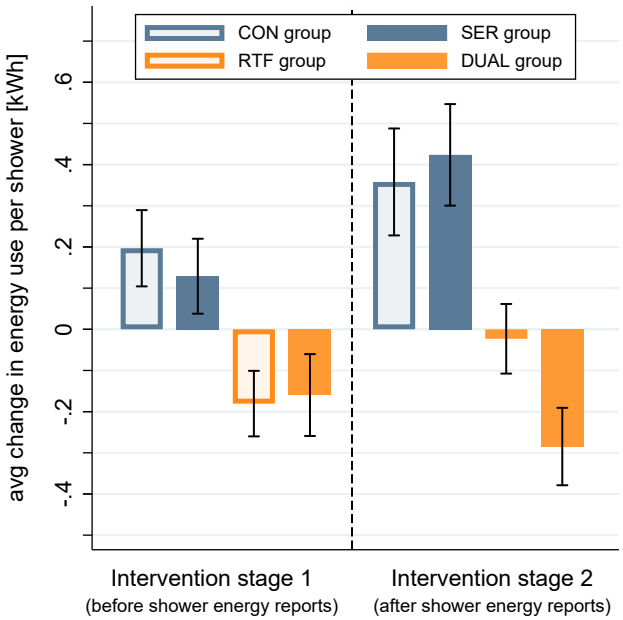


Figure 2.6.1. Descriptive evidence on energy conservation effects

Notes: The bars represent changes in average energy use per shower compared to the baseline period. The error whiskers show standard errors of the mean. Non-uploaders in SER and DUAL as well as non-uploaders without technical problems in CON and RTF are excluded.

2.6 Empirical results

2.6.1 Main results

First, we present descriptive evidence on the conservation effects of our interventions. Figure 2.6.1 shows subjects' average changes in energy 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. In order to show the TOT for 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. The graph essentially summarizes our main results in eight bars.

The four bars to the left of the dashed vertical line represent the change in energy 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 energy consumption drastically, by almost 0.4 kWh 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 that average energy 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).[30] 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 at around 0.4 kWh per shower. The third observation is that providing shower energy reports in isolation does not seem to result in effective behavioral change: energy 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.

Our descriptive results presented in Figure 2.6.1 are confirmed by formal empirical estimates based on the empirical strategy outlined in the previous section. We first focus on estimating the effect of real-time feedback in isolation, before turning to the effect of shower energy reports, for which we need to account for imperfect compliance.

The cleanest way to estimate the effect of real-time feedback is to only compare subjects in the RTF and CON groups over the entire intervention period, by estimating equation (2.5.1). Table 2.6.1 columns 1 and 2 show that real-time feedback in isolation reduces resource use by 0.40 kWh of energy and 6.3 liters of water per shower compared to the CON group, which corresponds to about 17-18% of baseline use. Columns 3 and 4 present the results from estimating equation (2.5.2) on the full sample, using treatment assignment as the independent variable. Subjects in RTF and DUAL conserved about 0.31 kWh of energy and 4.6 liters of water per shower in intervention stage 1, compared to subjects in CON and SER. These are slightly lower than the estimates in columns 1 and 2, partly due to the inclusion of the DUAL and SER groups, partly due to the conservation effect increasing in intervention stage 2 (albeit statistically insignificantly).

Result 1. *Real-time feedback through the smart meter display led to a reduction in energy (water) consumption by around 0.3-0.4 kWh (4.6-6.3 liters) or 14-18% per shower.*

With the advent of shower energy reports in intervention stage 2, we split the pairs up into the four separate groups again, which incidentally gives us ITT estimates for the effect of shower energy reports; but as discussed earlier, this misses the policy-relevant effect of actually receiving information through shower energy

30. 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.

Table 2.6.1. Effect of real-time feedback and ITT estimates

	only RTF & CON		*Intention to treat*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
Intervention	0.283***	4.453***	0.179***	2.915***
	(0.104)	(1.597)	(0.067)	(1.049)
Intervention × RTF/DUAL	-0.397***	-6.346***	-0.309***	-4.628***
	(0.125)	(1.926)	(0.087)	(1.387)
IN stage 2			0.187*	3.157**
			(0.097)	(1.441)
IN stage 2 × RTF/DUAL			-0.071	-1.745
			(0.118)	(1.854)
IN stage 2 × SER			0.038	0.147
			(0.130)	(2.006)
IN stage 2 × DUAL			-0.133	-2.302
			(0.093)	(1.555)
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 2.A.4.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

reports. The ITT estimates for the effect of shower energy reports are neither significant for SER nor DUAL, but the point estimates for the DUAL group look quantitatively relevant.

Therefore, we move on to the TOT analyses described in Section 2.5 to estimate the effect of actually receiving information through shower energy reports on conservation behavior. In Table 2.6.2, columns 1 and 2 show the estimates obtained by using the uploader-only approach, in which we estimate regression equation (2.5.3) on the restricted sample that excludes non-uploaders in SER and DUAL, as well as non-uploaders in RTF and CON without technical issues. Columns 3 and 4 display the LATE estimates, for which we use random treatment assignment to the SER or DUAL group as instruments for actually uploading data and receiving informative shower energy reports. 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 actual take-up (i.e. uploading data) is as good as random conditional on being willing to upload data.

Both approaches produce nearly identical results, suggesting that endogeneity of treatment take-up is not a major issue. The conservation effect of real-time feedback

Table 2.6.2. Treatment on the treated (TOT) estimates

	Uploaders-only		*LATE*	
	(1) Energy [kWh]	(2) Water [liter]	(3) Energy [kWh]	(4) Water [liter]
Intervention	0.179	2.628	0.172*	2.533
	(0.111)	(1.702)	(0.102)	(1.565)
Intervention × RTF/DUAL	-0.388***	-5.753***	-0.365***	-5.481***
	(0.134)	(2.124)	(0.125)	(1.981)
Intervention × SER	0.027	0.837	0.016	0.733
	(0.154)	(2.415)	(0.134)	(2.082)
Intervention × DUAL	0.035	0.576	0.109	2.159
	(0.113)	(1.860)	(0.107)	(1.751)
IN stage 2	0.150	2.770*	0.189*	3.273**
	(0.093)	(1.422)	(0.098)	(1.460)
IN stage 2 × RTF/DUAL	-0.021	-1.142	-0.053	-1.463
	(0.118)	(1.913)	(0.120)	(1.908)
IN stage 2 × SER	0.090	0.714	0.042	-0.084
	(0.137)	(2.168)	(0.162)	(2.510)
IN stage 2 × DUAL	-0.222**	-3.702**	-0.215*	-3.836*
	(0.100)	(1.756)	(0.116)	(2.037)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Clusters	261	261	318	318
Observations	14712	14712	17942	17942
R^2	0.413	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 for the main coefficients of interest is depicted in Figure 2.A.5.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

in isolation is also similar to the ones reported in Table 2.6.1. The results show that, contrary to our prediction, shower energy reports had no significant conservation effect in the SER group, and the point estimates even run in the opposite direction. While the null effect is not very tightly estimated, we can rule out energy use reductions of more than 7.5% per shower with 90% confidence in the (less precise) LATE specification. Furthermore, we can reject the hypothesis that shower energy reports in isolation were as effective as real-time feedback in isolation ($p < 0.003$ in all specifications).

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

Statistical imprecision aside, this does not imply that the shower energy reports are generally ineffective in our setting, but only when administered in isolation to the SER group. In stark contrast, we find that subjects in the DUAL group further reduced energy use by around 0.22 kWh (water use by around 3.8 liters) per shower in intervention stage 2, which corresponds to another 10 percentage points reduction from baseline consumption. This means that adding shower energy reports boosted the effectiveness of real-time feedback by more than 50%. The difference between energy conservation effects in the DUAL group and the SER group is weakly significant in the uploaders-only specification ($p = 0.067$). Unfortunately, we do not have enough power to detect this differential effect with larger statistical certainty, due to the technical issues with the smartphone app. Our results are, however, fully robust to randomization-based inference methods (Young, 2019), as presented in Appendix Figures 2.A.4 and 2.A.5.

Result 3. *Combining real-time feedback with shower energy reports further reduced energy (water) use by around* 0.22 *kWh (3.8 liters) per shower and thus boosted the conservation effect of real-time feedback in isolation by more than* 50%.

This contrast between the effect of shower energy reports with and without real-time feedback is all the more remarkable given that subjects in DUAL had already cut their energy consumption per shower significantly in response to real-time feedback and thus had less room for further behavioral adjustments, which is exactly one of the opposing effects against complementarity we described in the theoretical framework. Overall, there seems to be a strong complementarity between real-time feedback and shower energy reports. This is consistent with our theoretical framework, which shows that in the presence of multiple sources of bias to resource conservation, behavioral interventions may need to overcome 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, the lack of salience in resource consumption is likely to hinder conservation efforts. Real-time feedback through smart meters could thus turn environmental considerations into action by putting them into focus while showering. We will analyze the underlying mechanisms more closely in Section 2.7.

2.6.2 Treatment effect dynamics

We now investigate whether the conservation effects of real-time feedback and shower energy reports remain stable over the three-month period of our study. The previous subsection already documents that the effect of real-time feedback does not drop from the first to the second intervention stage. Therefore, we now focus on the 5-6 weeks period of IN stage 2. To estimate dynamic effects, we extend the empirical model for average treatment effects (equation 2.5.3) by interacting with a time variable Z_i:

Table 2.6.3. Treatment effect dynamics

	$Z_i = \mathbb{I}\{post\ 2nd\ report\}$		$Z_i = \#\ weeks\ after\ 1st\ report$	
	(1) Uploaders	(2) LATE	(3) Uploaders	(4) LATE
IN stage 2	0.139	0.176	0.065	0.109
	(0.103)	(0.110)	(0.124)	(0.127)
IN stage 2 × RTF/DUAL	-0.027	-0.053	0.047	0.019
	(0.128)	(0.134)	(0.156)	(0.159)
IN stage 2 × SER	0.092	0.048	0.198	0.174
	(0.148)	(0.169)	(0.181)	(0.202)
IN stage 2 × DUAL	-0.068	-0.041	0.030	0.075
	(0.123)	(0.135)	(0.166)	(0.177)
IN stage 2 $\times Z_i$	0.019	0.022	0.032	0.029
	(0.093)	(0.090)	(0.027)	(0.026)
IN stage 2 × RTF/DUAL $\times Z_i$	0.012	0.000	-0.026	-0.026
	(0.123)	(0.119)	(0.037)	(0.035)
IN stage 2 × SER $\times Z_i$	-0.002	-0.010	-0.041	-0.051
	(0.126)	(0.136)	(0.042)	(0.047)
IN stage 2 × DUAL $\times Z_i$	-0.279	-0.316	-0.099	-0.114*
	(0.209)	(0.215)	(0.064)	(0.067)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Clusters	261	318	261	318
Observations	14712	17942	14712	17942
R^2	0.413	0.005	0.413	0.005

The results are obtained by estimating equation (2.6.1). The full table with all the coefficients is presented in Appendix A Table 2.A.3. 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.
$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

$$
\begin{aligned}
y_{it} \;=\; & \alpha_i + IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D} + \beta_2 T_i^{H} + \beta_3 T_i^{D}\right) \\
& + IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^{H} + \gamma_3 T_i^{D}\right) \\
& + IN_{it}^{s2} \times Z_i \times \left(\delta_0 + \delta_1 T_i^{R/D} + \delta_2 T_i^{H} + \delta_3 T_i^{D}\right) + \varepsilon_{it}\,. \qquad (2.6.1)
\end{aligned}
$$

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 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 2.6.3 shows that the effect of shower energy reports in the DUAL group seems to gradually unfold over time. In fact, the reduction in energy use is not yet statistically significant in the first two weeks of intervention stage 2; columns (1) and (2) show that the average conservation effect is driven largely by the final 3-4 weeks of the study, i.e. 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 0.1 kWh every week. We should note these changes over time are mostly statistically insignificant and therefore to be interpreted with caution. Shower energy reports in isolation (SER group) show no signs of any dynamic pattern; the coefficient is identical before and after the second report. 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 months of our experiment.[31]

There are several potential explanations for the pattern of increasing behavioral responses over time that we observe in the DUAL group. For one, subjects may have skimmed through the email reports initially and only looked at it more carefully later. What speaks against this explanation is that most of the subjects responded to the attached mini-surveys within few days after we sent out the email and that the overall response rate was much higher in the first than in the second report.[32] Nevertheless, it may well be possible that the apparent increase over time is at least partly due to lower measurement error of when subjects where actually treated. Also, the social comparison in the second report might have provided additional motivation, which then interacted with real-time feedback, as would be predicted by the theoretical framework. A third explanation is that subjects may have required some time to try out and discover new strategies for further reducing energy use. This experimentation channel seems consistent with the finding that subjects in the DUAL group do not conserve energy by reducing their shower duration in the second intervention stage, but rather through adjusting flow rate and water temperature. 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).

2.6.3 Heterogeneous treatment effects

Particular subgroups of individuals may have responded more strongly to our interventions than others. Previous studies often find that households or individuals with high baseline consumption tend to respond more strongly to policy interven-

31. This is consistent with other studies using the Amphiro smart meter. For example, Agarwal, Fang, Goette, et al. (2020) find stable effects for an intervention duration of up to 6 months, as well as evidence for strong persistence several months after the intervention.

32. In fact, 53% (40%) of all subjects in DUAL responded within one day of receiving the first (second) report, and 80% (48%) did so within one week. Overall response rate was 81% (48%).

tions targeted at their conservation behavior (e.g. Allcott (2011), Ferraro and Price (2013), and Tiefenbeck, Goette, et al. (2018)). For example, Allcott (2011) reports that Opower home energy reports achieved virtually no savings for households in the bottom decile of baseline energy use, whereas the treatment effect for top-decile users was 6.3% savings. Tiefenbeck, Goette, et al. (2018) estimate that real-time feedback has an additional conservation effect of 0.31 kWh for a 1 kWh increase in baseline energy use per shower. Policy makers concerned about cost-effectiveness can therefore purposefully target high-baseline users.

To estimate heterogeneity along the dimension of baseline energy use, we extend the basic statistical model in equation (2.5.3) with interaction terms:

$$\begin{aligned} y_{it} \;=\; & \alpha_i + IN_{it} \times \left(\beta_0 + \beta_1 T_i^{R/D} + \beta_2 T_i^{H} + \beta_3 T_i^{D}\right) \\ & + IN_{it} \times X_i \times \left(\lambda_0 + \lambda_1 T_i^{R/D} + \lambda_2 T_i^{H} + \lambda_3 T_i^{D}\right) \\ & + IN_{it}^{s2} \times \left(\gamma_0 + \gamma_1 T_i^{R/D} + \gamma_2 T_i^{H} + \gamma_3 T_i^{D}\right) \\ & + IN_{it}^{s2} \times X_i \times \left(\mu_0 + \mu_1 T_i^{R/D} + \mu_2 T_i^{H} + \mu_3 T_i^{D}\right) + \varepsilon_{it} \end{aligned} \tag{2.6.2}$$

where variable X_i is a measure of subjects' baseline energy consumption per shower. As a measure of baseline consumption, we use a subject's average energy use in the 9 baseline showers (the first shower is excluded), re-centered around the sample mean (2.21 kWh) so that intercept terms can be interpreted as effects at the mean. In addition, we report a specification where X_i is an above-median indicator.

Table 2.6.4 presents TOT estimates of heterogeneous effects along baseline energy use. Note that we only show the main coefficients of interests here to keep the table visually tractable, but the full set of coefficients can be found in Table 2.A.4 in the Appendix. Consistent with previous literature, we find that the effect of real-time feedback in isolation increases with baseline use. In intervention stage 2, compounding the effects over both periods ($\hat{\lambda}_1 + \hat{\mu}_1$), subjects with 1 kWh higher baseline reduce their energy use per shower by an additional 0.26 kWh ($p = 0.069$) on average. Above-median baseline users (mean 3.30 kWh) save 0.63 kWh ($p = 0.039$) of energy more per shower compared to subjects with below-median baseline use (mean 1.17 kWh). This is consistent with the notion that real-time feedback reduces "slack" in resource use, but does not lead subjects to compromise on basic needs. It also appears that providing information through shower energy reports in the DUAL condition induces about double the conservation effect for above-median users ($\hat{\gamma}_3 + \hat{\mu}_3 = -0.322$ kWh, $p = 0.075$), compared to below-median baseline users ($\hat{\gamma}_3 = -0.156$ kWh, $p = 0.096$) in intervention stage 2, although the difference is not significant ($p = 0.414$). Shower energy reports in isolation (SER group) are neither effective for low- nor high-baseline users. In fact, it seems that subjects with below-median baseline use tend to increase their energy use in intervention stage 2 ($p = 0.028$).

Table 2.6.4. Treatment effect heterogeneity

	(1) continuous	(2) $\mathbb{I}\{> median\}$
...	...	...
Intervention × RTF/DUAL	-0.403***	-0.254***
	(0.127)	(0.096)
IN stage 2 × RTF/DUAL	-0.014	0.171*
	(0.117)	(0.102)
IN stage 2 × SER	0.095	0.267**
	(0.139)	(0.121)
IN stage 2 × DUAL	-0.239**	-0.156*
	(0.102)	(0.093)
...	...	...
Intervention × RTF/DUAL × Baseline energy use	-0.164	-0.247
	(0.119)	(0.266)
IN stage 2 × RTF/DUAL × Baseline energy use	-0.094	-0.385*
	(0.101)	(0.228)
IN stage 2 × SER × Baseline energy use	-0.021	-0.368
	(0.124)	(0.268)
IN stage 2 × DUAL × Baseline energy use	-0.097	-0.166
	(0.092)	(0.203)
Other treatment variables	*yes*	*yes*
Individual fixed effects	*yes*	*yes*
Observations	14675	14675
R^2	0.413	0.413

The coefficients are obtained by estimating equation (2.6.2). For visual ease, not all coefficient estimates are presented. The full table with is can be found in Appendix A Table 2.A.4. 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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

2.7 Underlying mechanisms

The empirical results show that, in our setting, shower energy reports seem to be ineffective in isolation, but induce large and significant conservation effects when combined with real-time feedback, which suggests that our interventions are strong complements. Through the lens of the theoretical framework in section 2.2, the most plausible mechanism for this finding is that the two interventions operated through complementary policy levers. 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 face bias due to 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 policy levers, we would generally not expect complementarities unless there is 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 main empirical results.

2.7.1 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, with low-baseline users overestimating and high-baseline users underestimating their water use (see Figure 2.4.1). The picture changes completely after the interventions. Figure 2.7.1 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 2.A.5 is presented in Appendix A. Whereas subjects in the CON group remain as ignorant as before, subjects who received real-time feedback (RTF and DUAL group) are now able to estimate their water use almost without bias, so the fitted regression lines are close to the identity line. While the slope looks slightly flatter for the DUAL group compared to the RTF group, the difference is not statistically significant. Importantly, shower energy reports in isolation (SER group) also induce strong learning effects about water use, as estimated water use increases visibly in actual water use per shower (TOT slope 0.57), and significantly more strongly than in the CON group ($p = 0.025$). We cannot reject that learning through shower energy reports is more effective with real-time feedback than without ($p = 0.497$). While these analyses focus on the bias of subjects' estimates (conditional on actual water use), we obtain similar results when we look at the magnitude of absolute estimation errors across groups. Table 2.A.6 in Appendix A shows that subjects in the three treated groups are on average about 27-30 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 did engage with the interventions and thereby became more aware of their own water use behavior in

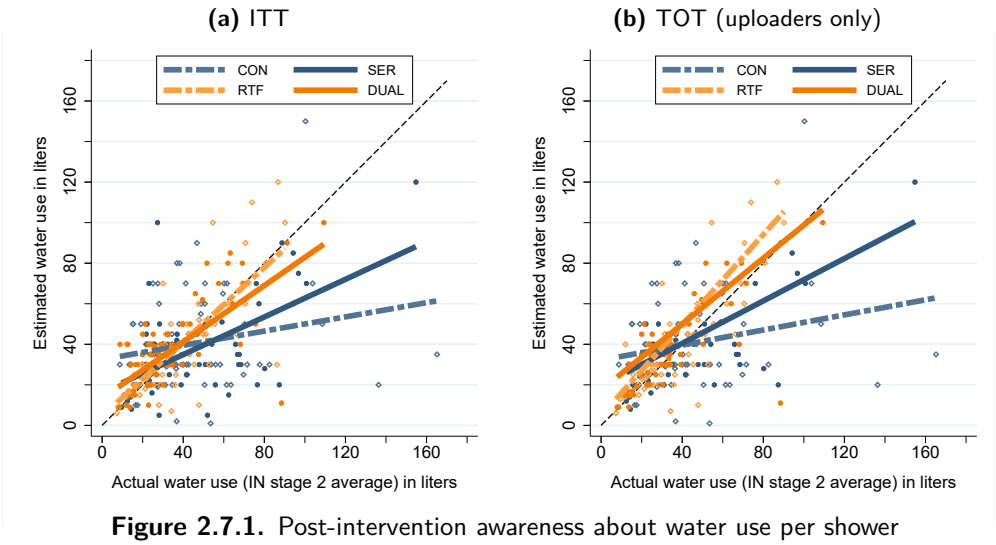


Figure 2.7.1. Post-intervention awareness about water use per shower

Notes: Both graphs compare subject's water use estimates in the final questionnaire with their actual water use in intervention stage 2. 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.

the shower. However, belief updates about water use per se are unlikely to drive our main results. First, subjects' prior beliefs about water use were by and large unbiased on average. Second, although the posterior beliefs in the SER group do not become quite as accurate as in the RTF group, we would have expected at least some conservation effect through shower energy reports in isolation if belief updating about water use was the main mechanisms. 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 use, but also on energy use 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 2.4.6). 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 the SER sheet induces surveyees to drastically adjust their beliefs about CO_2 emissions upwards com-

pared to the CON or RTF sheets ($p < 0.001$). This experimentally-induced belief update is further associated with a 0.24 standard deviations ($p = 0.003$) increase in self-stated intention to take shorter showers in the future, compared to the RTF sheet group. For further details, see Appendix E.

Shower energy reports seem to induce knowledge gains about the environmental impact of showering, yet they are only associated with significant conservation effects when combined with real-time feedback. One of the key insights of our theoretical framework is that if multiple sources of bias play a role, different behavioral interventions can become complements, because a single narrowly-targeted intervention is undermined by the presence of other behavioral biases. Hence, our empirical results suggest that, in the absence of real-time feedback, additional barriers like limited attention or self-control problems have prevented knowledge gains and good intentions from translating into actual behavior.

2.7.2 Engagement with shower energy reports

One potential alternative channel is differential treatment engagement, in the sense that subjects in different treatment groups may pay 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 as described in the theoretical framework. The previous subsection shows that shower energy reports did induce significant learning effects about water use in the shower also in the SER group. Furthermore, we can also directly assess whether the level of scrutiny was similar in the SER and the DUAL group. To do so, we make use of the mini-surveys that were attached to each of the two report emails. As described before, each email included a link to a survey in which we asked subjects to give an estimate of the amount of water they use in a typical 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.

Table 2.A.7 in Appendix A shows response rates by treatment group in the uploaders-only sample. Recall that subjects in the RTF and CON groups received Placebo emails containing a link to the same mini-survey. The overall response rate of uploaders was 87% for the first email and 71% for the second email. 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. Apart from the extensive margin, Table 2.A.7 further shows subjects' relative estimation error by treatment group, defined as percent deviation of estimated water use in the mini-survey from the actual water use

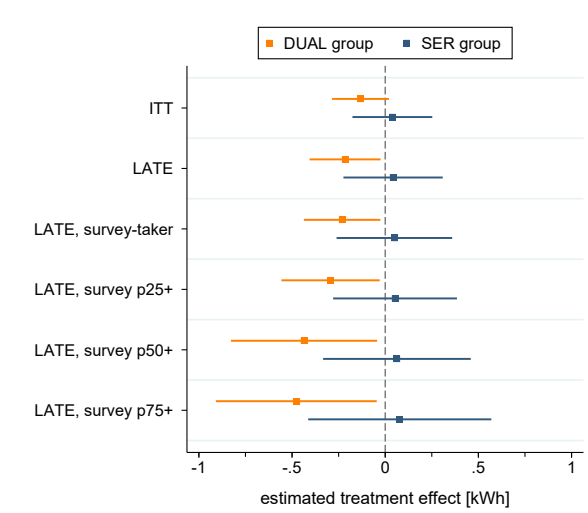


Figure 2.7.2. Effects for different levels of engagement with shower energy reports

Notes: The points 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 the 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, was above the 25th, 50th, or 75th percentile of all subjects, respectively.

per shower.[33] Smaller estimation errors are an indication of subjects paying closer attention while reading the reports. Respondents in the SER group were only 10% off on average, and they actually gave more precise estimates than respondents in the DUAL group ($p = 0.039$), who were 21% off on average. Notwithstanding, both groups still outperform the CON group (49% off on average) by far. Overall, we find no evidence that uploaders in the DUAL group studied reports more carefully than uploaders in SER group.

As an additional plausibility check that it is not lower level of engagement with the shower energy reports that prevented energy conservation in the SER group, we look at whether subjects who studied the reports more closely also engaged more strongly in conservation actions. For this purpose, we again make use of subjects' water use assessments in the mini-surveys and regress energy use per shower on several new shower energy report treatment indicators that increase in their level 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 subject's estimate precision, defined as distance between estimated and measured water use per shower, was above the 25th, 50th, or 75th percentile of all subjects, respectively. To avoid the endogeneity issue at hand, we use treatment assignment as

33. As measure for actual water use per shower, we take the number that was calculated for each subject when sending out the shower energy reports.

instrument for level of engagement with reports. Figure 2.7.2 plots the coefficients and confidence intervals for the effect of shower energy reports in SER and DUAL group, respectively. The estimated conservation effect in the DUAL group increases monotonically with the strictness of our compliance definition, reaching almost 0.5 kWh for the strictest indicator. In contrast, even the most studious subjects in the SER group did not reduce their energy use in response to the reports, which corroborates our interpretation that some source of bias such as limited attention may have prevented shower energy reports in isolation from inducing behavioral change.

2.7.3 Other potential mechanisms

There are a number of alternative channels through which our interventions could affect conservation behavior. For one, they could trigger Hawthorne effects, but recall that also subjects in the control group received a smart meter and placebo emails reminding them to upload their data. See Appendix 2.F for a more detailed discussion of why Hawthorne or cueing effects are unlikely to explain our findings. Another potential channel may be that we made the activity of showering less enjoyable to our subjects. However, our survey results indicate that subjective shower comfort was not affected by real-time feedback or shower energy reports, thus also alleviating concerns about unintended negative welfare effects of the interventions (Allcott and Kessler, 2019). Furthermore, we find no evidence for increases in subjects' general pro-environmental attitude. If anything, we observe a decrease in self-perceived pro-environmental attitudes in the treated groups compared to the control group, potentially due to feedback provision curbing the capacity for distorted self-image formation. We report these results based on survey data in Appendix 2.F.

2.8 Conclusion

In this paper, we argued that if multiple sources of behavioral bias (e.g., imperfect information and 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 theoretical framework that delineates the interplay of behavioral interventions and illustrates mechanisms for complementarity and substitutability in a setting with multiple behavioral biases; in particular, the potential for complementarity becomes larger the more differentiated the interventions are with regard to their targeted biases. We then presented results from a three-month field experiment on energy conservation behavior in a specific resource-intensive everyday activity (showering), in which we evaluated interaction effects between two types of interventions: shower energy reports, which provided information on energy use and carbon emissions via email, and real-time feedback through a smart

meter display, which made water consumption in the shower immediately salient. While only the latter induced a significant conservation effect when implemented in isolation, combining both interventions resulted in a striking complementarity. It seems that knowledge gains about environmental impacts only triggered conservation behavior once resource use was additionally made salient through real-time feedback.

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 0.22 kWh (3.8 liters). For comparison, total daily energy use for lighting in German households is about 0.33 kWh per person on average.[34] In his influential evaluation of the Opower home energy reports, which target *aggregate* electricity use in U.S. households, Allcott (2011) finds a household-level conservation effect of 0.62 kWh per day. One limitation of our study is that we do not observe subjects' consumption behavior outside the shower. However, in a related study that uses Amphiro smart shower meters in a representative household sample in Singapore, Schmitt, Tiefenbeck, Fang, et al. (2021) find that the direct conservation effect in the shower may even understate the effect on overall household water consumption. This is in line with recent evidence for potential positive spillover effects of pro-environmental interventions (Jessoe, Lade, Loge, et al., 2021; Sherif, 2021).

We attempted to make a step towards understanding why different interventions can be complements (or substitutes). While both our theoretical framework and our field experiment are tailored to a very specific setting, the notion that potentially multiple different barriers need to be overcome for behavioral change can be relevant in other contexts as well, including situations involving more standard economic barriers such as lack of incentives or constraints on time, money, or technology. Such complexity of behavioral mechanisms is a pervasive feature of many domains of our lives, and it is likely that this creates numerous opportunities for complementarities between different interventions, yet many of these may still be untapped.[35]

34. Source: German Federal Statistical Office.

35. Indeed, some empirical findings in the literature are at least suggestive of mechanisms at work that are similar to the one we suggest. For example, Cortes, Fricke, Loeb, et al. (2019) find that text-message based curricula supporting good parenting practices work less well when parents face high cognitive load than during time periods when the load is lighter. Dupas and Robinson (2013) study financial savings behavior in a developing country and find that simply providing a safe box for storing money is already quite 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 short-run temptations, increases the effectiveness of healthy purchasing subsidies (Brownback, Imas, and Kuhn, 2019).

Further research is necessary to investigate whether this channel for complementarity of interventions we propose also generalizes to other settings and more representative samples. Nevertheless, our study underlines that any evaluation is inevitably confined to the particular policy and choice environment that consumers act in, which may itself be malleable. Interventions that may seem feeble at first glance may thus be able to unfold their full potential once combined with other interventions that address remaining sources of behavioral bias. For example, our results suggest a special role for interventions that increase the salience of one's resource use: giving individuals simple tools that allow them to track their use may also make their behavior more sensitive to other policies, such as price incentives (Jessoe and Rapson, 2014). Hence, behavioral policy design should not only consider through which channels a particular intervention affects behavior, but also attempt to identify and overcome behavioral barriers that may still remain.

New policies are always introduced to an existing set of policies, institutions, and norms. As social scientists are beginning to pioneer the process from small-scale proof-of-concept studies to large-scale interventions (Banerjee, Banerji, Berry, et al., 2017), future research should therefore synchronously advance our knowledge on the interplay of different policy instruments.

Appendix 2.A Supplementary figures and tables

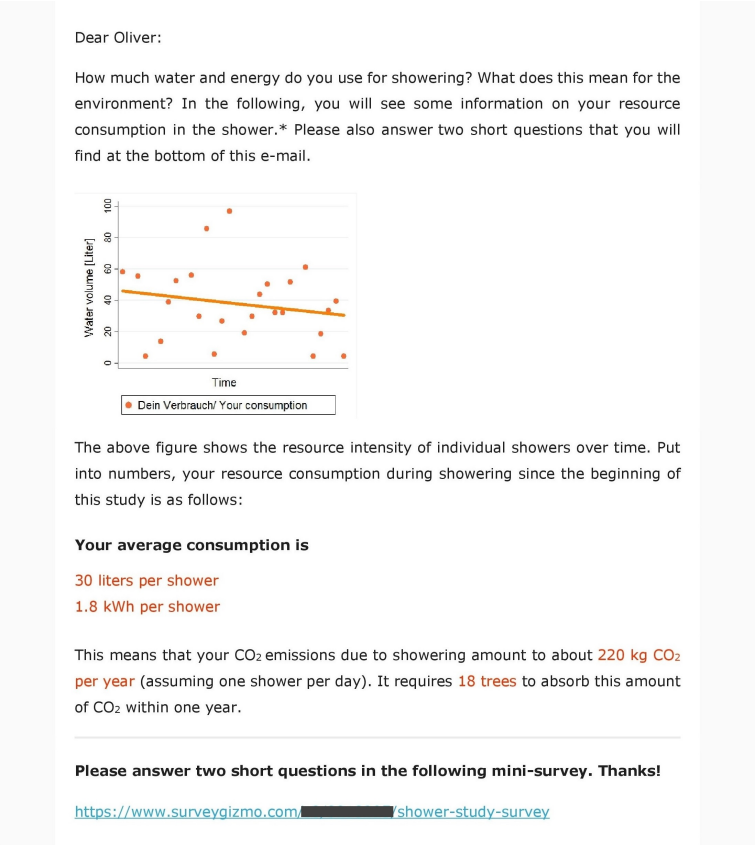


Figure 2.A.1. Screenshot of a typical shower energy report (for a fictitious person)

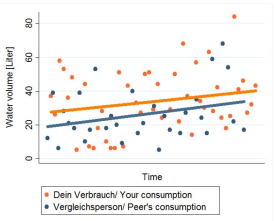
universitätbonn

Universität zu Köln

Feedback on your resource use in the shower

Dear Oliver:

How much water and energy do you use for showering? What does this mean for the environment? In the following, you will see some information on your resource consumption during in the shower.* Furthermore, you will see the consumption behavior of another (anonymous) study participants, who was randomly assigned to you. Please also answer two short questions that you will find at the bottom of this e-mail.



The above figure shows the resource intensity of your and your peer's showers over time. To put this in numbers, resource consumption during showering since the beginning of this study is as follows:

Your average consumption is	**Your peer's avg. consumption is**
34 liters per shower	26 liters per shower
2 kWh per shower	1.5 kWh per shower

This means that your CO_2 emissions per year due to showering amount to about 250 kg CO_2 (assuming one shower per day). It requires 20 trees to absorb this amount of CO_2 within one year.

Figure 2.A.2. Screenshot of a shower energy report with peer comparison

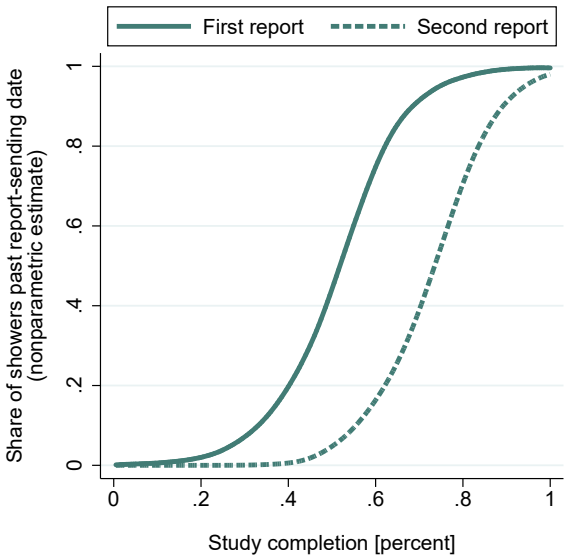


Figure 2.A.3. Empirical distribution of report timing

Table 2.A.1. 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 2.A.2. Comparing uploaders and non-uploaders

	uploaders: mean (sd)	non-uploaders: mean (sd)	diff. in means *p*-value
Energy [kWh]	2.23 (1.38)	2.20 (1.37)	0.95
Water volume [liter]	38.54 (22.36)	37.13 (20.73)	0.87
Temperature [Celsius]	35.41 (3.33)	35.94 (3.47)	0.61
Flow rate [liter/min]	6.01 (2.34)	5.30 (2.19)	0.11
Duration [min]	6.61 (2.98)	7.69 (4.54)	0.10
Environmental attitude	-0.04 (1.03)	0.07 (0.93)	0.79
Shower comfort	-0.05 (1.05)	0.14 (0.87)	0.55
1 if female	0.58 (0.49)	0.70 (0.46)	0.28
Age in years	23.93 (3.80)	23.79 (3.99)	0.95
1 if international	0.31 (0.46)	0.41 (0.49)	0.42
Observations	228	90	

Subject characteristics before sending out shower energy reports. *p*-values adjusted for multiple hypothesis testing (Romano-Wolf procedure using 2,000 bootstrap repetitions).

Table 2.A.3. Treatment effect dynamics

	$Z_i = \mathbb{I}\{post\ 2nd\ report\}$		$Z_i = \#\ weeks\ after\ 1st\ report$	
	(1) Uploaders	(2) LATE	(3) Uploaders	(4) LATE
Intervention	0.179	0.172*	0.178	0.171*
	(0.111)	(0.103)	(0.111)	(0.102)
Intervention × RTF/DUAL	-0.388***	-0.365***	-0.386***	-0.364***
	(0.134)	(0.125)	(0.133)	(0.125)
Intervention × SER	0.027	0.016	0.029	0.019
	(0.154)	(0.134)	(0.154)	(0.134)
Intervention × DUAL	0.046	0.119	0.047	0.120
	(0.113)	(0.108)	(0.112)	(0.108)
IN stage 2	0.139	0.176	0.065	0.109
	(0.103)	(0.110)	(0.124)	(0.127)
IN stage 2 × RTF/DUAL	-0.027	-0.053	0.047	0.019
	(0.128)	(0.134)	(0.156)	(0.159)
IN stage 2 × SER	0.092	0.048	0.198	0.174
	(0.148)	(0.169)	(0.181)	(0.202)
IN stage 2 × DUAL	-0.068	-0.041	0.030	0.075
	(0.123)	(0.135)	(0.166)	(0.177)
IN stage 2 $\times Z_i$	0.019	0.022	0.032	0.029
	(0.093)	(0.090)	(0.027)	(0.026)
IN stage 2 × RTF/DUAL $\times Z_i$	0.012	0.000	-0.026	-0.026
	(0.123)	(0.119)	(0.037)	(0.035)
IN stage 2 × SER $\times Z_i$	-0.002	-0.010	-0.041	-0.051
	(0.126)	(0.136)	(0.042)	(0.047)
IN stage 2 × DUAL $\times Z_i$	-0.279	-0.316	-0.099	-0.114*
	(0.209)	(0.215)	(0.064)	(0.067)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Clusters	261	318	261	318
Observations	14712	17942	14712	17942
R^2	0.413	0.005	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 2.A.4. Treatment effect heterogeneity

	X_i : *baseline energy use*		X_i : *envir. attitude*	
	(1) linear	(2) median$^+$	(3) linear	(4) median$^+$
Intervention	0.180*	0.272***	0.178	0.243
	(0.105)	(0.072)	(0.112)	(0.203)
Intervention × RTF/DUAL	-0.403***	-0.254***	-0.392***	-0.324
	(0.127)	(0.096)	(0.134)	(0.222)
Intervention × SER	0.012	-0.139	0.003	0.030
	(0.146)	(0.112)	(0.149)	(0.255)
Intervention × DUAL	0.085	0.020	0.049	-0.088
	(0.111)	(0.087)	(0.113)	(0.151)
IN stage 2	0.148	0.001	0.166*	0.140
	(0.091)	(0.065)	(0.089)	(0.172)
IN stage 2 × RTF/DUAL	-0.014	0.171*	-0.036	-0.032
	(0.117)	(0.102)	(0.115)	(0.196)
IN stage 2 × SER	0.095	0.267**	0.074	0.214
	(0.139)	(0.121)	(0.133)	(0.221)
IN stage 2 × DUAL	-0.239**	-0.156*	-0.225**	-0.313**
	(0.102)	(0.093)	(0.105)	(0.157)
Intervention $\times X_i$	-0.016	-0.192	0.031	-0.137
	(0.101)	(0.226)	(0.130)	(0.220)
Intervention × RTF/DUAL $\times X_i$	-0.164	-0.247	-0.210	-0.176
	(0.119)	(0.266)	(0.145)	(0.269)
Intervention × SER $\times X_i$	0.109	0.325	-0.172	-0.039
	(0.140)	(0.301)	(0.166)	(0.296)
Intervention × DUAL $\times X_i$	0.062	0.039	0.103	0.310
	(0.110)	(0.215)	(0.105)	(0.232)
IN stage 2 $\times X_i$	0.056	0.313*	-0.076	0.056
	(0.077)	(0.179)	(0.116)	(0.185)
IN stage 2 × RTF/DUAL $\times X_i$	-0.094	-0.385*	0.084	-0.002
	(0.101)	(0.228)	(0.129)	(0.237)
IN stage 2 × SER $\times X_i$	-0.021	-0.368	0.083	-0.363
	(0.124)	(0.268)	(0.144)	(0.260)
IN stage 2 × DUAL $\times X_i$	-0.097	-0.166	0.024	0.146
	(0.092)	(0.203)	(0.083)	(0.207)
Individual fixed effects	*yes*	*yes*	*yes*	*yes*
Clusters	260	260	257	257
Observations	14675	14675	14501	14501
R^2	0.413	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 2.A.5. Estimated vs actual water use per shower

	before study	after study	
		ITT	TOT
Actual volume	0.271	0.175	0.186
	(0.263)	(0.139)	(0.145)
Actual volume × RTF	0.025	0.742***	0.835***
	(0.376)	(0.199)	(0.179)
Actual volume × SER	-0.465	0.289*	0.381**
	(0.292)	(0.174)	(0.169)
Actual volume × DUAL	-0.074	0.520***	0.517**
	(0.299)	(0.182)	(0.230)
RTF group	-0.131	1.694	3.162
	(6.777)	(3.234)	(3.183)
SER group	-7.001	-4.578	-5.200*
	(5.813)	(3.181)	(3.029)
DUAL group	-5.182	1.655	1.588
	(5.851)	(3.136)	(3.826)
Constant	43.436***	39.507***	39.610***
	(4.590)	(2.429)	(2.542)
Observations	267	296	251
R^2	0.030	0.378	0.440

Robust standard errors in parentheses. Actual volume is recentered around 40 liters.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.6. Estimated versus actual water use: relative estimation error

	before study	after study	
		ITT	TOT
RTF group	0.075	-0.283***	-0.296***
	(0.201)	(0.073)	(0.075)
SER group	0.008	-0.172**	-0.281***
	(0.175)	(0.080)	(0.072)
DUAL group	-0.055	-0.214**	-0.270***
	(0.178)	(0.085)	(0.076)
Constant	0.927***	0.577***	0.583***
	(0.136)	(0.061)	(0.064)
Observations	302	296	251
R^2	0.002	0.050	0.101

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

Table 2.A.7. Response to mini-surveys attached to reports

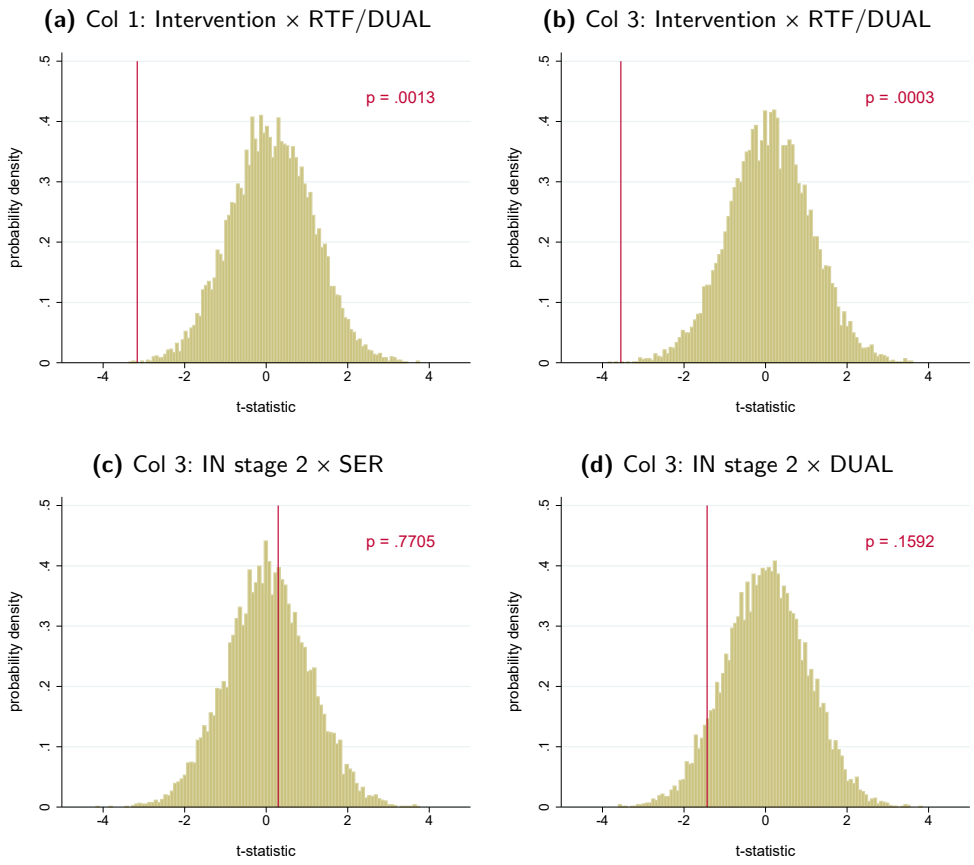
	Survey response rate			
	(1) first report	(2) second report	(3) any report	(4) estimation error [%p]
RTF group	-1.05	0.53	-2.48	-30.69
	(5.35)	(6.57)	(4.90)	(7.62)
SER group	-7.85	-16.76	-7.18	-38.74
	(6.39)	(7.91)	(5.81)	(7.54)
DUAL group	0.58	-26.17	-0.44	-27.70
	(5.54)	(8.14)	(5.00)	(8.57)
Constant	88.89	80.56	91.67	48.93
	(3.73)	(4.70)	(3.28)	(7.14)
p-value for SER = DUAL	0.203	0.308	0.270	0.039
Observations	261	261	261	231
R-squared	0.009	0.061	0.008	0.139

Robust standard errors in parentheses.

Table 2.A.8. Margins of behavioral adjustment

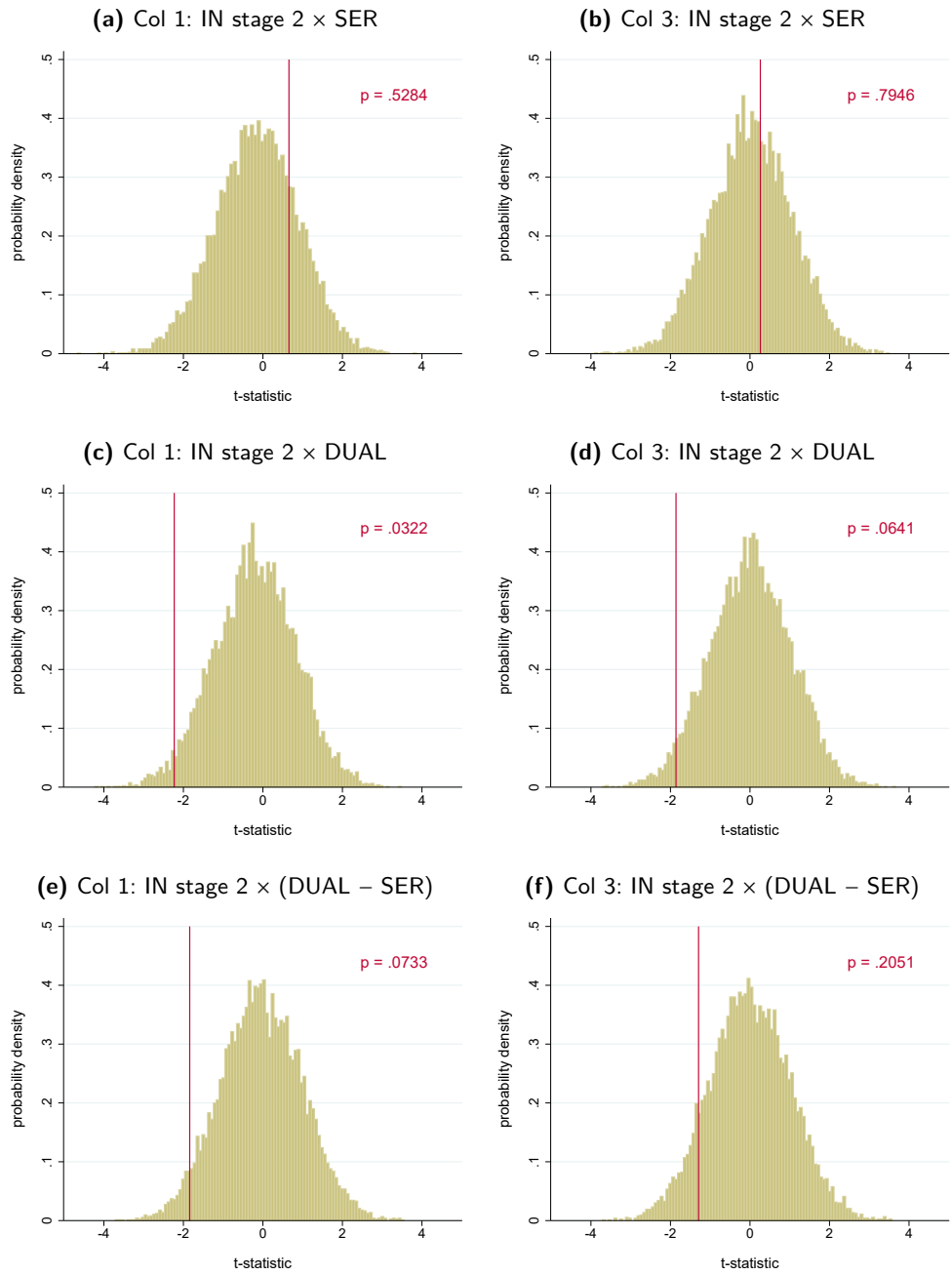
	Duration in seconds			Temperature in °C			Flow rate in liter/min		
	(1) ITT	(2) Uploaders	(3) LATE	(4) ITT	(5) Uploaders	(6) LATE	(7) ITT	(8) Uploaders	(9) LATE
Intervention	8.40	8.52	8.40	-0.00	0.01	-0.00	0.25**	0.25**	0.25**
	(11.02)	(11.76)	(11.02)	(0.34)	(0.37)	(0.34)	(0.11)	(0.12)	(0.11)
Intervention × RTF/DUAL	-38.41**	-39.32**	-38.41**	-0.74	-0.88*	-0.74	-0.17	-0.19	-0.17
	(16.31)	(17.17)	(16.31)	(0.46)	(0.49)	(0.46)	(0.17)	(0.18)	(0.17)
Intervention × SER	13.37	11.16	11.50	-0.54	-0.63	-0.53	-0.08	-0.12	-0.07
	(16.35)	(17.60)	(16.19)	(0.42)	(0.46)	(0.41)	(0.15)	(0.18)	(0.15)
Intervention × DUAL	6.35	7.73	6.32	-0.23	0.19	-0.25	-0.01	-0.20	-0.04
	(16.92)	(17.49)	(16.55)	(0.40)	(0.41)	(0.38)	(0.18)	(0.19)	(0.17)
IN stage 2	24.69	12.99	24.69	0.39	0.46	0.39	0.26**	0.24**	0.26**
	(17.75)	(9.81)	(17.75)	(0.31)	(0.34)	(0.31)	(0.11)	(0.12)	(0.11)
IN stage 2 × RTF/DUAL	-32.30	-19.75	-32.30	0.28	0.31	0.28	0.21	0.22	0.21
	(19.74)	(13.33)	(19.74)	(0.40)	(0.42)	(0.40)	(0.19)	(0.20)	(0.19)
IN stage 2 × SER	-30.85	-12.45	-38.01	0.18	0.21	0.22	0.16	0.11	0.20
	(21.83)	(17.21)	(27.21)	(0.37)	(0.40)	(0.46)	(0.17)	(0.17)	(0.21)
IN stage 2 × DUAL	-0.49	0.71	-0.62	-0.21	-0.41	-0.26	-0.34	-0.40	-0.43
	(12.70)	(14.22)	(16.01)	(0.35)	(0.33)	(0.44)	(0.22)	(0.25)	(0.28)
Observations	17942	14712	17942	17942	14712	17942	17942	14712	17942
R^2	0.383	0.361	0.001	0.310	0.323	0.003	0.751	0.763	0.016

Standard errors in parentheses (clustered on subject level)
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$



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.

Figure 2.A.4. Randomization inference for coefficients of interest in Table 2.6.1



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.

Figure 2.A.5. Randomization inference for coefficients of interest in Table 2.6.2

Appendix 2.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 2.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, Goette, 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 2.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

$$\widetilde{IN}_{it}^{s2} = \begin{cases} IN_{it}^{s2} & \text{if uploader} \\ \widehat{Pr}\left(IN_{it}^{s2} = 1 \mid \pi, uploader\right) & \text{if non-uploader}. \end{cases} \tag{2.D.1}$$

Appendix 2.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 person's 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 2.E.1 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

Table 2.E.1. 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.

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 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 2.F More on Other Potential Mechanisms

2.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 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 2.7.2 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.

2.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 2.F.1 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 2.F.1 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 underly-

Table 2.F.1. 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.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

ing 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.

References

Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter. 2005. "A Review of Intervention Studies Aimed at Household Energy Conservation." *Journal of Environmental Psychology* 25 (3): 273–291. [63]

Agarwal, Sumit, Ximeng Fang, Lorenz Goette, Samuel Schoeb, Thorsten Staake, Verena Tiefenbeck, and Davin Wang. 2020. "The Role of Goals in Motivating Behavior: Evidence from a Large-Scale Field Experiment on Resource Conservation." *mimeo*, [92]

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9-10): 1082–1095. [60, 63, 93, 100]

Allcott, Hunt. 2016. "Paternalism and Energy Efficiency: An Overview." *Annual Review of Economics* 8 (1): 145–176. [59]

Allcott, Hunt, and Judd B. Kessler. 2019. "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons." *American Economic Journal: Applied Economics* 11 (1): 236–276. [99]

Allcott, Hunt, and Todd Rogers. 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10): 3003–3037. [76, 92]

Andor, Mark, Andreas Gerster, Joerg Peters, and Christoph M. Schmidt. 2020. "Social Norms and Energy Conservation Beyond the US." *Journal of Environmental Economics and Management* 103: 102351. [63]

Andor, Mark A., and Katja M. Fels. 2018. "Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects." *Ecological Economics* 148: 178–210. [63]

Ashraf, Nava, B. Kelsey Jack, and Emir Kamenica. 2013. "Information and Subsidies: Complements or Substitutes?" *Journal of Economic Behavior & Organization* 88: 133–139. [64]

Attari, Shahzeen Z. 2014. "Perceptions of Water Use." *Proceedings of the National Academy of Sciences of the United States of America* 111 (14): 5129–5134. [59]

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi 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 (37): 16054–16059. [59, 65, 76, 81]

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31 (4): 73–102. [101]

Banerjee, Abhijit, Arun Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew Jackson, Harini Kannan, Francine Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha. 2021. "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization." *Working Paper*, [64]

Brandon, Alec, John A. List, Robert D. Metcalfe, Michael K. Price, and Florian 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 (12): 5293–5298. [64]

Brownback, Andy, Alex Imas, and Michael Kuhn. 2019. "Behavioral Food Subsidies." *Working Paper*, [100]

Byrne, David P., Andrea La Nauze, and Leslie A. Martin. 2018. "Tell Me Something I Don't Already Know: Informedness and the Impact of Information Programs." *Review of Economics and Statistics* 100 (3): 510–527. [61, 82]

Camilleri, Adrian R., Richard P. Larrick, Shajuti Hossain, and Dalia Patino-Echeverri. 2019. "Consumers Underestimate the Emissions Associated with Food but are Aided by Labels." *Nature Climate Change* 9 (1): 53–58. [59]

Carlsson, Fredrik, Christina Annette Gravert, Verena Kurz, and Olof Johansson-Stenman. 2021. "The Use of Green Nudges as an Environmental Policy Instrument." *Review of Environmental Economics and Policy* 15 (2): 216–237. [63]

Coe, David T., and Dennis J. Snower. 1997. "Policy Complementarities: The Case for Fundamental Labor Market Reform." *IMF Staff Papers* 44 (1): [60]

Cortes, Kalena E., Hans Fricke, Susanna Loeb, David S. Song, and Ben York. 2019. "When Behavioral Barriers Are Too High or Low: How Timing Matters for Parenting Interventions." *IZA Discussion Paper No. 12416*, [100]

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio. 2013. "Information Strategies and Energy Conservation Behavior: A Meta-analysis of Experimental Studies from 1975 to 2012." *Energy Policy* 61: 729–739. [63]

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "Education, HIV, and Early Fertility: Experimental Evidence from Kenya." *American Economic Review* 105 (9): 2757–2797. [63]

Dupas, Pascaline, Elise Huillery, and Juliette Seban. 2018. "Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon." *Journal of Economic Behavior & Organization* 145: 151–175. [64]

Dupas, Pascaline, and Jonathan Robinson. 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review* 103 (4): 1138–1171. [63, 100]

Fanghella, Valeria, Matteo Ploner, and Massimo 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. [64]

Ferraro, Paul J., and Michael K. Price. 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95 (1): 64–73. [93]

Fischer, Corinna. 2008. "Feedback on Household Electricity Consumption: A Tool for Saving Energy?" *Energy Efficiency* 1 (1): 79–104. [63]

Frederiks, Elisha R., Karen Stenner, and Elizabeth V. Hobman. 2015. "Household Energy Use: Applying Behavioural Economics to Understand Consumer Decision-Making and Behaviour." *Renewable and Sustainable Energy Reviews* 41: 1385–1394. [59]

Gabaix, Xavier. 2017. "Behavioral Inattention." *NBER Working Papers 24096*, [67]

Gardner, Gerald T., and Paul C. Stern. 2008. "The Short List: The Most Effective Actions U.S. Households Can Take to Curb Climate Change." *Environment and Behavior* 50 (5): 12–24. [60]

Gerster, Andreas, Mark Andor, and Lorenz Goette. 2020. "Disaggregate Consumption Feedback and Energy Conservation." *CEPR Discussion Paper 14952*, [63]

Giaccherini, Matilde, David H. Herberich, David Jimenez-Gomez, John A. List, Giovanni Ponti, and Michael K. Price. 2020. "Are Economics and Psychology Complements in Household Technology Diffusion? Evidence from a Natural Field Experiment." *Working Paper*, [64]

Hahn, Robert, Robert D. Metcalfe, David Novgorodsky, and Michael K. Price. 2016. "The Behavioralist as Policy Designer: The Need to Test Multiple Treatment to Meet Multiple Targets." *NBER Working Paper 22886*, [64]

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning Through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics* 129 (3): 1311–1353. [67]

Holladay, J. Scott, Jacob LaRiviere, David Novgorodsky, and Michael Price. 2019. "Prices versus nudges: What matters for search versus purchase of energy investments?" *Journal of Public Economics* 172: 151–173. [64]

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467. [85]

Jamison, Julian C., Dean Karlan, and Jonathan Zinman. 2014. "Financial Education and Access to Savings Accounts: Complements or Substitutes? Evidence from Ugandan Youth Clubs." *NBER Working Paper 20135*, [63]

Jessoe, Katrina, Gabriel E. Lade, Frank Loge, and Edward Spang. 2021. "Spillovers from Behavioral Interventions: Experimental Evidence from Water and Energy Use." *Journal of the Association of Environmental and Resource Economists* 8 (2): 315–346. [100]

Jessoe, Katrina, and David Rapson. 2014. "Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use." *American Economic Review* 104 (4): 1417–1438. [63, 101]

Karlin, Beth, Joanne F. Zinger, and Rebecca Ford. 2015. "The Effects of Feedback on Energy Conservation: A Meta-analysis." *Psychological Science* 141 (6): 1205–1227. [63]

Kollmuss, Anja, and Julian Agyeman. 2002. "Mind the Gap: Why Do People Act Environmentally and What Are the Barriers to Pro-Environmental Behavior?" *Environmental Education Research* 8 (3): 239–260. [59]

List, John A., Robert D. Metcalfe, Michael K. Price, and Florian Rundhammer. 2017. "Harnessing Policy Complementarities to Conserve Energy: Evidence from a Natural Field Experiment." *NBER Working Paper 23355*, [64]

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, And Complementarities In Education: Experimental Evidence From Tanzania." *Quarterly Journal of Economics* 134 (3): 1627–1673. [63, 64]

Myers, Erica, and Mateus Souza. 2019. "Social Comparison Nudges Without Monetary Incentives: Evidence from Home Energy Reports." *E2e Working Paper 041*, [63]

Nolan, Jessica M., P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2008. "Normative Social Influence is Underdetected." *Personality and Social Psychology Bulletin* 34 (7): 913–923. [78]

Schmitt, Kathrin, Verena Tiefenbeck, Ximeng Fang, Lorenz Goette, Thorsten Staake, and Davin Wang. 2021. "Pro-environmental spillover effects in the resource conservation domain: Evidence from a randomized controlled trial in Singapore." *mimeo*, [100]

Schwartz, Daniel, and George Loewenstein. 2017. "The Chill of the Moment: Emotions and Proenvironmental Behavior." *Journal of Public Policy & Marketing* 36 (2): 255–268. [76, 92]

Sherif, Raisa. 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*, [100]

Tiefenbeck, Verena. 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*, [87]

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake. 2018. "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation." *Management Science* 64 (3): 1458–1476. [59–61, 63, 65, 75–77, 93, 113]

Tiefenbeck, Verena, Anselma Woerner, Samuel Schoeb, Elgar Fleisch, and Thorsten Staake. 2019. "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives." *Nature Energy* 4: 35–41. [63]

Tolstoy, Leo. 2003. *Anna Karenina*. (First published in Russian, 1873-1877; translation by Richard Pevear and Larissa Volokhonsky). London: Penguin Books. [66]

Tonke, Sebastian. 2019. "Imperfect Knowledge, Information Provision and Behavior: Evidence from a Field Experiment to Encourage Resource Conservation." *Working Paper*, [60]

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *Quarterly Journal of Economics* 134 (2): 557–598. [90]

Chapter 3

Motivating the Adoption of Digital Contact Tracing Apps in the Covid-19 Pandemic

Joint with Lorenz Goette and Zihua Chen

3.1 Introduction

One important tool in the controlling of the Covid-19 pandemic lies in fast and effective detection of individuals who have been in close contact with infected persons. Test-trace-and-isolate strategies are thus an important tool for slowing and containing the spread of the virus by breaking infection chains (Kretzschmar, Rozhnova, Bootsma, et al., 2020; Contreras, Dehning, Loidolt, et al., 2021; Fetzer and Graeber, 2021). Digital contact tracing (DCT) apps can complement manual contact tracing by automatically tracking potential contact persons and sending out notifications after a positive test result is enters. This can reduce time lags between discovery of an infection and the identification and informing of contact persons, thus reducing the risk of cascading infection chains due to tracing delay. At least 40 countries in the world have developed and launched DCT apps (Ahmed, Michelin, Xue, et al., 2020; O'Neill, Ryan-Mosley, and Johnson, 2020). However, the effectiveness of DCT apps depends crucially on the adoption rate, with some studies simulating that far more than the majority of the population needs to adopt the app for it to have a sufficient impact on containing the pandemic (Ferretti, Wymant, Kendall, et al., 2020; Pollmann, Pollmann, Wiesinger, et al., 2020), although empirical evidence from suggests that already lower adoption rates can help in slowing the spread of the SARS-CoV-2 (Wymant, Ferretti, Tsallis, et al., 2021). Thus, installation and usage of DCT apps constitutes a public good, as the benefits accrue to society at large, beyond just the individual who uses it.

Unfortunately, adoption rates in most countries are relatively low. For example, at the time of our study (October 2020), the German Corona-Warn-App (henceforth CWA) had only been downloaded about 20 million times since its launch on June 16, 2020, which corresponds to less than 25% of the total population in Germany, with DCT apps in other countries face similar problems (Mosoff, Friedlich, Scassa, et al., 2020). As installing the DCT app is voluntary in most countries, it is important to find effective communication and promotion strategies to encourage higher adoption rates in the population, thus bearing resemblance to the challenges we might face in convincing voluntary uptake of vaccination against COVID-19.[1]

In this paper, we conduct a large-scale randomized online experiment in Germany ($N \approx 1$ million) to test whether providing individuals with information about current Covid-19 infection rates in their place of residence can encourage adoption of DCT apps. We build on a standard digital advertisement video, a twelve-second clip encouraging adoption of the app, commissioned by the German government and deployed on social media platforms such as Facebook or Instagram, by appending an additional video element to the beginning of the ad that provides individuals with feedback about the current seven-day incidence rate (i.e. Covid-19 infections over the last seven days per 100,000 inhabitants) in their county of residence.[2] Furthermore, we add a regional comparison element by showing whether (and by how much) the county's incidence is above or below their state's average incidence rate.

In our two main experimental conditions, we compare the effectiveness of the regular promotion video (T0) with the treatment video that additionally provides feedback on local incidence rates and regional comparisons (T1). As one might be concerned that when incidence rates are low, highlighting this aspect (and highlighting below-average status) may lead to a boomerang effect, we introduce a second treatment condition T2, which adds on T1 by further displaying normative messages that appeals to users in below-average counties to maintain the status quo, and to users in above-average counties to challenge the status quo. The full experimental design is summarized in Figure 3.1.1. As mentioned, the regular promotion video has a duration of 12 seconds and basically presents three animated frames that encourage viewers to install the app. Our basic treatment video (T1) starts with a 6-second animated frame that provides feedback on how the regional incidence rate in the county of residence compares to that of other counties in the same state; it also includes a smiley or frowny, respectively, to amplify the normative impact of the feedback (Schultz, Nolan, Cialdini, et al., 2007). Afterwards, the video segues into the regular promotion video for the CWA. Detailed information on the

1. At the time of first writing (February 2021), Germany was still at the beginning of its vaccination campaign.

2. We use official data on Covid-19 incidence rates obtained from the Robert-Koch-Institut (RKI), the federal government agency and research institute responsible for disease control and prevention in Germany.

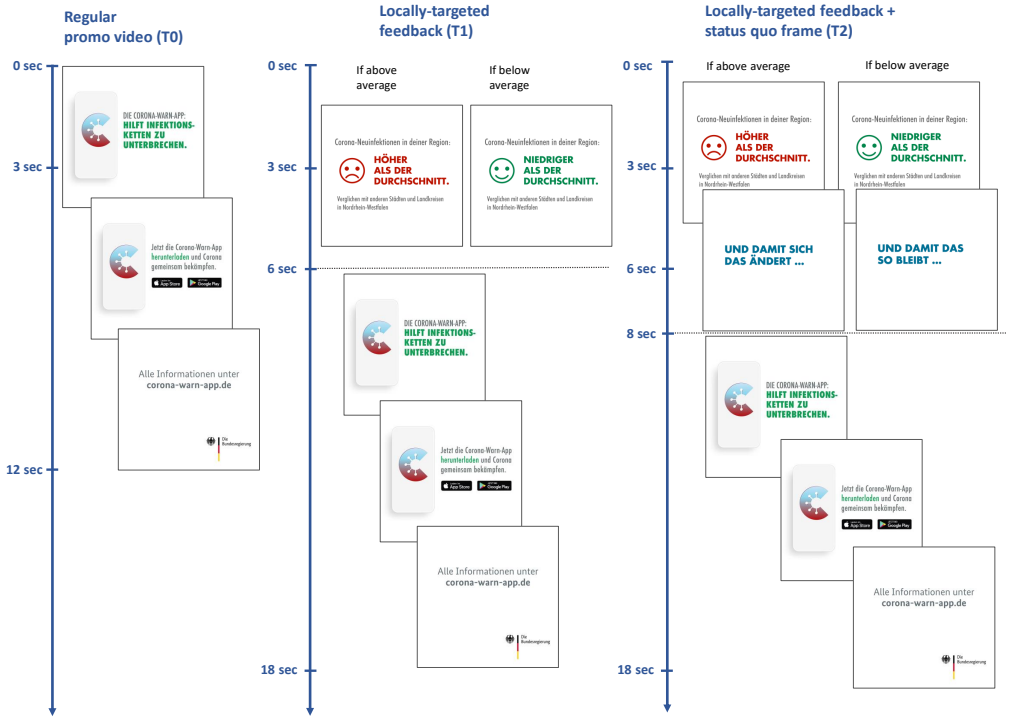


Figure 3.1.1. Depiction of video ads in different experimental conditions

Notes: Depiction of the video ad treatments used in our online intervention. The figure shows which frames were displayed in the respective videos at a given play time. In T1 the treatment screen stayed on for 6 seconds at the beginning and then continued with the standard promotion video. In T2, the first treatment screen time was also 6 seconds, but it was followed by a normative message frame for 2 seconds before starting the standard video. The display duration of the very last frame was reduced by 2 seconds in T2 to keep the overall video duration of T1 and T2 comparable at about 18 seconds (compared to 12 seconds for T0). All frames were animated.

7-day-incidence is included in a description text that accompanies the ad. The T2 ad further enhances the ad by the status quo message in a 3-second frame after the regional comparison.

Highlighting local incidence rates can affect the motivation to install the DCT app through several behavioral channels. Firstly, it may make the impact of CWA adoption on local Covid-19 infections more salient and tangible, similar to feedback interventions in other domains (e.g., Ferraro and Price, 2013; Tiefenbeck, Goette, Degen, et al., 2018), thus potentially reducing the perceived social distance of its public benefits (Small and Loewenstein, 2003). Secondly, comparing one's home county with neighboring counties can prime an individual's identification with his or her county of residence. Group identification has been shown to increase prosocial contributions to the ingroup in lab experimental settings (e.g. Charness, Rigotti, and Rustichini, 2007; Chen and Li, 2009; Goette, Huffman, Meier, et al., 2012; Böhm and Rockenbach, 2013; Charness and Holder, 2019), but few studies have at-

tempted to explicitly make use of group comparisons based on region of residence to motivate public good contributions in the field (Kessler and Milkman, 2018). Thirdly, in the T2 treatment, we reinforce the group comparisons with normative loss-framed messages appealing to status quo considerations, which may further increase motivation (Kahneman and Tversky, 1979; Schultz et al., 2007). Finally, it may also be the case that the information screen on local incidence rates simply draws in more viewers at the beginning and thereby increases the number of users who are exposed to the CWA promotion; this is related to the "stopping power" of marketing and the importance of capturing and retaining consumers' attention with advertisement (Drèze and Hussherr, 2003; Pieters and Wedel, 2004; Pieters, Wedel, and Bartra, 2010; Texeira, Wedel, and Pieters, 2012).

We test our interventions in a large-scale online experiment on social media between October 7 to October 17, 2020. Using Facebook Ads, we target users based on their county of residence and deliver regionally-customized video advertisements for the CWA on Facebook and Instagram. Using the A/B testing functionality, we randomly assign whether users are exposed to T0, T1, or T2. We focus our ads on 221 of the largest counties in Germany and reach a total of 1,115,404 different users. For all ads, we embed a link to the official CWA homepage, and we use click-through rate (CTR), i.e. the share of users who follow the link, as our main outcome variable, as following the link to the CWA website indicates at the minimum a strong interest in seeking more information and at best the intention to immediately download the app. As additional intermediary variables, we collect measures of video view rates to investigate user engagement with the video ads. In particular, we use 3-second view rates as proxy for initial attention in order to analyze potential extensive margin effects. To complement the experiment on social media, we conduct an online survey with a representative sample in Germany (N = 5,830) into which we embed the same interventions T0, T1, and T2. This allows us to better understand beliefs and attitudes towards the CWA as well as how they respond to information about local incidence rates. See *Materials and Methods* for more information on samples and study procedures.

At least two additional mechanisms might work against the positive effects from providing feedback about local incidence rates. First, it is possible that highlighting incidence rates triggers negative emotions in individuals, which may lead them to seek to avoid engaging with the topic (Golman, Hagmann, and Loewenstein, 2017; Golman, Loewenstein, Molnar, et al., 2020). Information avoidance could lead to a negative interaction between the treatment and local incidence rates. Second, it may also be that counties with high Covid-19 incidence rates have different demographic composition from low-incidence counties, and that same bundle of demographic characteristics that lead to higher Covid-19 incidence rates are independently also predictive of less concern for the local spread of Covid-19 or interest in the DCT app. Our empirical setup allows us to examine this, as we observe strong regional differences in Covid-19 incidence rates as well as strong changes over time. We predict

local incidence rates across the entire study period with county-level characteristics. This allows us to decompose local incidence rates into a baseline "demographic" incidence rate and a time-varying residual. We can then test which of the components determines regional heterogeneity in response to the CWA ads.

3.2 Background

The German Corona-Warn-App (CWA) was commissioned by the German Federal Government and jointly developed by Deutsche Telekom and SAP and is based on the digital privacy-preserving proximity tracing (DP-3T) app architecture. Once installed, the app can detect other app users in proximity using Bluetooth and exchange encrypted identification numbers (IDs) between devices. When an app user enters a positive Sars-CoV-2 test result into the app, all users who have been in proximity with that person will be automatically informed, but the identity of the contact remains anonymous. It has been available for download on Google Play Store for and Apple Store since 16 June 2020. Within the first month of deployment, the CWA was downloaded about 15 million times, but the uptake rate quickly flattened. By October 2020, when our experiment was conducted, the app had been downloaded 20 million times, which corresponds to less than 25% of the total population in Germany. At the time of writing this thesis (December 2021), the number of downloads is 34 million. However, these are upper bounds for the actual take-up rates of the CWA, since one person could have downloaded the app multiple time (e.g. on multiple devices) or deinstalled the app subsequently.

Note that at the time we conducted our study, the CWA only featured its core functionality of contact tracing. Over the course of 2021, several additional features have gradually been added, such as an integrated contact diary, information screens on Covid-19 incidence rates, hospitalization rates, vaccination rates, as well as an event check-in function and inclusion of a digital Covid-19 vaccination pass.

The pandemic situation in Germany at the time of our study was characterized by initially low but quickly increasing Covid-19 incidence rates, as well as relative leniency in government-mandated contact regulations. The nationwide 7-day Covid-19 incidence per 100,000 population was about 25 at the beginning of the study (Oct 7, 2020) and doubled to about 50 within the ten-day study period. This was accompanied by growing concern voiced by public health experts and government officials about the onset of a second wave in Germany, which proved to be correct. However, there were no significant tightening of restrictions until end of October, when a partial lockdown was announced. Thus, our interventions took place in a time period in which individuals were likely to have considerable interest in information about regional Covid-19 incidence rates and possible measures against the spread of the virus.

3.3 Empirical results

3.3.1 Effectiveness of the social media intervention

To examine the effectiveness of including locally targeted feedback on Covid-19 incidence rates for promoting use of the CWA, we compare the performance of our video ad treatments. Facebook Ads collects several performance metrics for video ads, including the share of users who have played the video for a at least a certain duration — 3 seconds, 25%, 50%, 100%, etc. — and the share of users who clicked on the embedded link to the CWA website (click-through rate). We use the latter as main outcome variable for our intervention, as following the link to the CWA website indicates at the minimum a strong interest in seeking more information and at best the intention to immediately download the app — note that the overwhelming share of impressions (96.43%) come from smartphone users. There might furthermore be positive effects on app adoption that are not captured by link clicks, since users might go directly to the app store to download the CWA without clicking on the link, or become more susceptible for installing the CWA in the future. As a necessary condition for any of these unmeasured effects is that users notice and pay some degree of attention to the CWA ad on their social media feed, we use 3-second video plays as an additional outcome variable that indicates the level of initial ad engagement. It is possible that the preview frames of the treatment videos that show feedback on the local incidence rate (see Figure 1) attract more attention than the regular preview frame of the control video, thus creating a window of opportunity to promote the CWA to more viewers. Unfortunately, all performance data is aggregated at the ad level, therefore we cannot jointly evaluate video play behavior and link clicks at the individual user level.

We make use of the random assignment of treatment videos to social media users to estimate the causal effect of local feedback on ad engagement as well as the likelihood to follow the link to the CWA website (click-through rate). The results for average video play rates and CTR are presented in Figure 3.3.1 and Appendix Table 3.B.1.

Figure 1a plots consumers' engagement with our ads using all information on video play rates at time marks that we receive from Facebook Ads. We first focus on 3-second video plays as an indicator of initial interest in the ad, i.e. the probability that a user stops and devotes attention to the video while browsing the Facebook or Instagram feed. About 7% of users play the treatment videos for at least 3 seconds, as compared to 5.3% in the control group, and the difference is highly statistically significant ($p < 0.001$). This 2.6%p effect is virtually identical for T1 and T2 — which was to be expected given that the two videos only differ after 6 seconds — and corresponds to a 30% increase relative to the control group. Next, we analyze 9-second video play, approximated using video plays at 75% for T0 and video plays at 50% for T1 and T2. This is an important mark, as users who watch the treatment

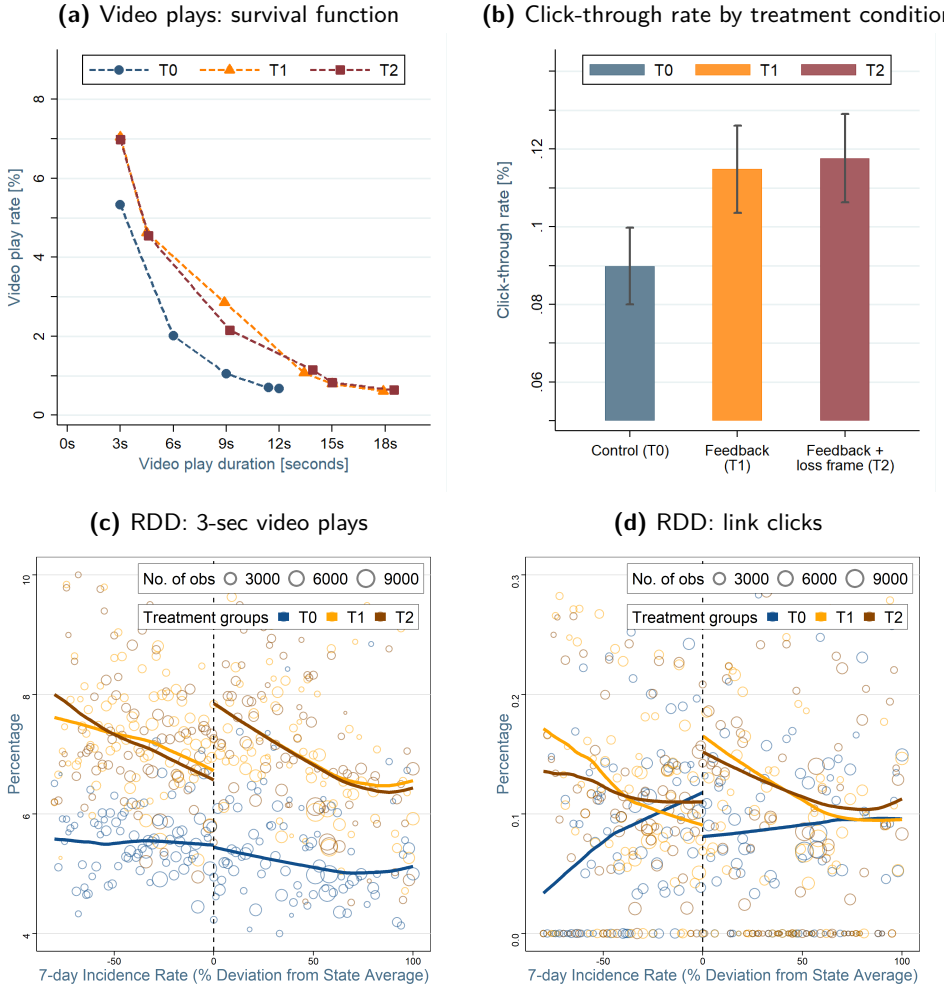


Figure 3.3.1. Effectiveness of locally targeted feedback interventions

Notes: Panel (a) shows the share of users who played the video for at least a certain duration. Panel (b) shows the average click-through rates in T0, T1, and T2, with error bars representing 95% confidence intervals. Panels (c) and (d) use a county-level regression discontinuity design (RDD) approach with local linear regressions (Triangular kernel, bandwidth = 20) to illustrate the effect of below- versus above-average incidence status.

videos for at least 9 seconds start being exposed to the regular CWA promotion ad. The share of users who continue to watch the video drops to 1.05% for the control ad, but remains significantly higher at 2.86% for T1 ($p < 0.001$) and 2.16% for T2 ($p < 0.001$). The discrepancy between T1 and T2 likely arises due to larger drop-out at the (static) status quo frame included in T2 but not in T1, and it is transitory, as the play rates start closely matching each other again at the next measurement point.[3] Tellingly, the share of users who completely watch the video from beginning

3. Furthermore, the exact video durations are 17.9 seconds for the T1 videos and 18.4 seconds for the T2 videos, thus our 9-second-view measure using 50% view rates is slightly imprecise and

to end is relatively similar for the control ads and the treatment ads, at around 0.64%, despite the latter being 50% longer in duration than the former. Overall, our locally targeted feedback interventions increase both initial as well as sustained attention to the social media advertisements, thus creating a window of opportunity for delivering public health recommendations such as DCT app adoption.[4]

Next, we investigate whether higher video play rates also manifest in higher propensity of users to follow the provided link in our experiment, our main outcome variable. Figure 1b shows that combining the CWA promotion video with locally targeted feedback on Covid-19 incidence increases the link click-through rate (CTR) by 0.25%p in T1 and 0.28%p in T2, corresponding to an effect size of around 30% relative to the control video. Appendix Table 3.B.1 shows that these effects are highly statistically significant at the one-percent level, but with no significant difference between T1 and T2 ($p = 0.396$). Thus, both our feedback interventions were effective in getting more people interested in the CWA. However, the overall prospects of our intervention become far more modest against the backdrop of the generally low baseline interest in the app, as only about 1 in 1000 users click on the embedded link to the CWA website. This is not so abysmal as it may seem, given the notoriously high number of competing stimuli when using social media. Indeed, the treatment videos T1 and T2 in our study only seem to perform slightly worse than the median video advertisement based on benchmark reports (Adstage, 2020; Wordstream, 2020).[5]

Intriguingly, the treatment effects of our feedback intervention both on 3 second video plays and click-through rates are about 30%. This implies that the entire effect of local feedback on link clicks can be predicted by higher initial user attention (the extensive margin), so on average we do not observe a large effect of local feedback on the probability of link click *conditional* on playing the video for at least 3 seconds (the intensive margin). Indeed, Appendix Table 3.B.1 shows that the ratio of link clicks to 3 second views for the treatment videos is indistinguishable from the 1.7% rate of the control video. One might be tempted to conclude that the entire treatment effect on CTR is driven by the extensive margin, i.e. whether users start paying attention

more favorable for T1 than for T2. However, these imprecisions probably only have a minor influence on our results.

4. A Nielsen study commissioned by Facebook finds that more than 40% of effects on ad recall and purchase intention happen in first 3 video seconds, and more than 70% in first 10 seconds (Facebook, 2015).

5. For example, one benchmark report by Adstage (2020) indicates that in the first quarter of 2020, the median share of *all* clicks per impression (CTR-all), i.e. link clicks (CTR) plus any other type of click interaction (expand ad, like, comment, share), was around 1.1% for ads on the Facebook News Feed and 0.22% for ads on the Instagram Feed. Similarly, according to Wordstream (2020), the average CTR-all of Facebook ads based was 0.89% in 2019, with large variation by topic of advertisement, ranging from 0.45% for Science to 1.68% for Pets & Animals. In comparison, the CTR-all of our T1 and T2 ads based on all clicks is 0.74% on Facebook and 0.08% on Instagram, whereas for T0 it is 0.21% on Facebook and 0.06% on Instagram. Note that the treatments are not perfectly comparable with regard to all clicks on Facebook, because T0 contains no text to expand.

to the ads at all However, the local feedback in T1 and T2 may also attract a pool of viewers that is has a different average latent propensity to download the CWA, because, unlike in the control ad, the treatment ads do not make it immediately salient that their true purpose is to promote the CWA. So, in principle there is a scale effect (how many people are draw in), a composition effect (who is draw in), and potentially an actual intensive margin effect (how much does the propensity to download increase after seeing the information). Although we cannot distinguish between these effects explicitly for lack of individual-level information, it is worth keeping in mind that this adverse selection effect possibly masks meaningful effects at the intensive margin.

The next question we address is whether the effects of our intervention are vary depending on the actual content of the feedback intervention on local incidence rates. One particularly prominent variation in our treatment videos that it either shows favorable comparison feedback (complemented by a smiley in green) in counties whose incidence rate lies below the state average, or unfavorable comparison feedback (complemented by a frowny in red) in counties whose above-average incidence. However, the comparison status is endogenous, since there might be observable and unobservable county characteristics that are correlated both with the local incidence rate as well as with residents' response to public health messages. In particular, counties in which residents are on average more skeptical of and less compliant with public health measures, including adoption of the CWA, will likely have higher incidence rates. Therefore, we estimate the causal effect of above- vs below-average status in a regression discontinuity design (RDD), exploiting the presence of counties whose incidence rates are so close to the state average that it is plausibly quasi-random whether they are above or below this threshold. Note that as we use the state average as reference threshold, we can further control for absolute incidence rate in the county, because there is substantial variation of average incidence rates across the 16 federated states in Germany.[6]

Figures 1c and 1d plot link clicks and 3-second video plays by deviation of counties' incidence rates from their respective state average, with the counties left (right) of the threshold at zero receiving favorable (unfavorable) comparison feedback. Following standard procedures for RDD estimates, we fit local linear regressions of ad performance on the running variable (deviation of incidence rate from state average) separately to the left and to the right of the threshold (Lee and Lemieux, 2010). The discontinuities at the threshold then constitute estimates of the causal effect of above- vs below-average status. Indeed, we observe sharp upward jumps in 3-second play rates and click-through rates of the treatment videos when switching from favorable to unfavorable feedback; these upward jumps are not present for the control video. Panel A of Table 3.3.1 presents estimates using an alternative

6. For the three city-states Berlin, Hamburg, and Bremen, we use the average incidence rate in the other two respective states as benchmark.

Table 3.3.1. Heterogeneity in treatment responses by 7-day incidence rate

Panel A: Actual 7-day incidence	3s plays	9s plays	CTR
Treated	2.22***	1.65***	0.04***
	(0.14)	(0.07)	(0.01)
Above State Average	-0.09	0.21**	0.01***
	(0.22)	(0.10)	(0.02)
7-day Incidence (cases per 1,000)	2.87***	0.75*	0.12*
	(0.66)	(0.40)	(0.07)
Treated × Above State Average	1.08***	0.06	0.04**
	(0.25)	(0.13)	(0.02)
Treated × 7-day Incidence (cases per 1,000)	-3.89***	-0.77***	-0.13***
	(0.60)	(0.29)	(0.04)
Panel B: Predicted 7-day Incidence	3s plays	9s plays	CTR
Treated	3.08***	2.14***	0.07***
	(0.19)	(0.10)	(0.02)
Above State Average	-0.20	0.14	0.01
	(0.22)	(0.09)	(0.02)
7-day Incidence (Residual)	0.53	-0.59**	0.04
	(0.54)	(0.33)	(0.06)
Treated × Above State Average	1.22***	0.16*	0.04**
	(0.24)	(0.12)	(0.02)
Treated × 7-day Incidence (Predicted)	-6.72***	-2.42***	-0.21***
	(0.74)	(0.37)	(0.06)
Treated × 7-day Incidence (Residual)	-0.16	1.32***	-0.01
	(0.53)	(0.33)	(0.05)
County Fixed Effects	*Yes*	*Yes*	*Yes*
Day Fixed Effects	*Yes*	*Yes*	*Yes*
Impressions	1044271	1044271	1044271
Counties	221	221	221

Standard errors in parentheses are clustered at the county level; obtained from bootstrapping with 10,000 simulations in Panel B. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

approach by controlling for county fixed effects, hence exploiting variations in incidence rates and the comparison status over time. Note that we pool the T1 and T2 conditions, as the results for both are very similar. Using this approach, we again document that unfavorable comparison feedback leads to $1.08\%p$ higher 3-second play rates ($p < 0.01$) and $0.04\%p$ higher CTRs ($p < 0.05$) in the treated groups T1 and T2 relative to the control group. Thus, it seems that social media users responded considerably more strongly to negative comparison feedback showing that local incidence rates are above the state average. Interestingly, we also observe that the treatment effects generally tend to follow a negative trend with regard to the local Covid-19 incidence rate.

The apparently negative association between incidence rate and the effectiveness of our intervention has adverse distributional effects, since one might argue that

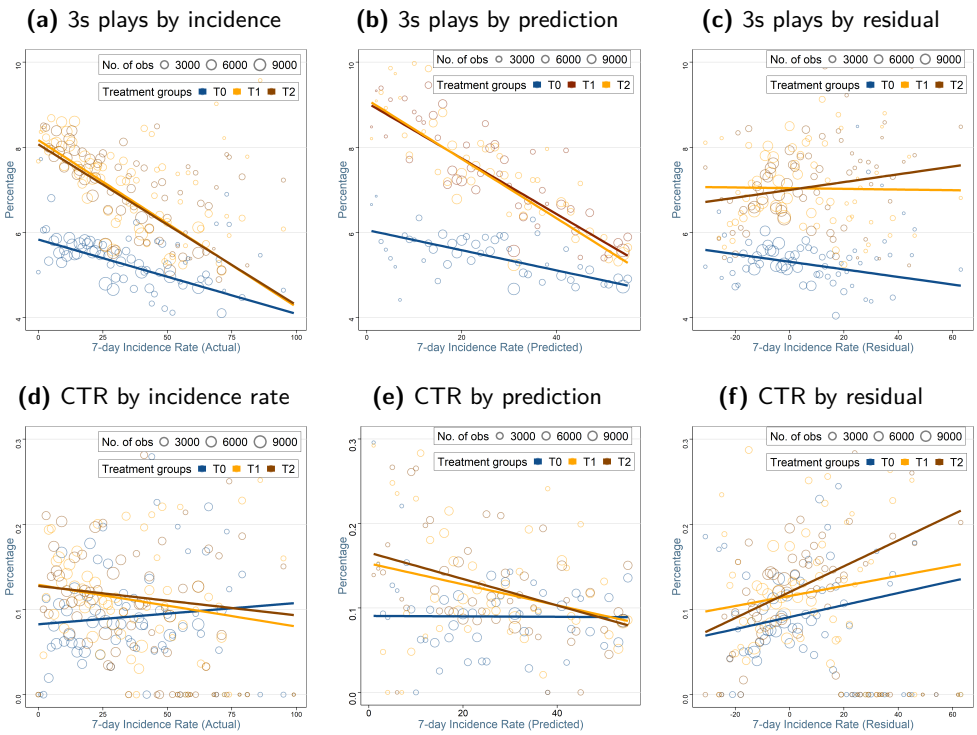


Figure 3.3.2. Heterogeneity by county-level incidence

Notes: All figures plot linear fits of 3 second video plays and click-through rates (CTR) by 7-day Covid-19 incidence rate in a county. Left: Actual incidence rate. Middle: Predicted incidence rates based on LASSO regressions on county characteristics (see Appendix Table 3.B.10). Right: Residual incidence rates.

the importance of fast and comprehensive contact tracing increases in proportion to the severity of the outbreak in a region. In Figure 3 (a) and (d), we plot linear regression estimates of ad performance against the absolute incidence rate in a county and confirm that both 3-second video plays and CTRs of the treatment videos generally decrease for counties with higher incidence rate. In fact, the average treatment effects documented before are predominantly driven by lower-incidence counties. While the number of link clicks for the control group video slightly increases with incidence rates, the 3-second plays show a particularly striking downward-sloping pattern both for the control video and, even more steeply, for the treatment videos. The formal regression estimates in Table 3.3.1 Panel A indicate that an increase in 7-day-incidence by 1 case per 1,000 population in the county decreases 3-second video plays by $3.89\%p$, 9-second video plays by $0.77\%p$, and CTRs by $0.13\%p$ for the treatment ads relative to the control ad, with all three interaction coefficients being highly statistically significant at the 1% level.

There can be two reasons for this large decrease in engagement with the treatment videos in high-incidence counties. First, users may have actively avoided paying attention to feedback on local incidence rate when it seemed to especially bad

news. While there is a burgeoning literature on the ostrich effect and information avoidance in general (e.g. Karlsson, Loewenstein, and Seppi, 2009; Golman, Hagmann, and Loewenstein, 2017; Golman, Loewenstein, et al., 2020), this explanation is at odds with the larger effect size for above-average status treatment videos. Second, it may simply be the case that certain characteristics of a county's population make it both more likely that incidence rate is high, e.g. due to lower compliance with social distancing and other recommended public health measures, and that average interest to the treatment ads is low.[7] To further examine this question, we merge our experimental data with supplementary data on various social, economic, and demographic characteristics of each county, obtained from the German statistical agencies. More specifically, we collect information on, among others, population density, age and gender profile, voting outcomes, migrant shares, economic indicators such as unemployment rates, etc., and use these time-invariant variables to predict each county's' incidence rates at the time of our study. Using LASSO with 10-fold cross-validation, we can explain about 50% of the variance in incidence rates across counties at the time of our study in mid-October.[8] We then proceed to decompose the incidence rate into two components: (1) the component that we can predict using various observable county characteristics, and (2) the residual component that comes from unobserved factors as well as sheer randomness.

Figure 3.3.2 (b), (c), (e), and (f) plot the 3-second video play rates and the CTRs separately by predicted and residual incidence rate. Interestingly, we find that the negative interaction effect of locally targeted feedback with the incidence rate is entirely driven by the predicted component, whereas treatment effects are, if anything, slightly increasing in the residual component. Table 1 Panel B presents the formal regression results. To adjust for the additional statistical uncertainty due to the predicted incidence rate being a generated regressor, we conduct inference using a bootstrap procedure (with 10,000 simulations) that also includes the LASSO prediction stage when estimating the second-stage standard errors. For a 1 per 1000 population increase in the *predicted* 7-day-incidence rate, the estimated treatment effect decreases by 6.72% for 3-second plays, by 2.42% for 9-second plays, and by 0.21%p for link click-through rates ($p < 0.001$ for all three coefficients). In stark contrast, the interaction with *residual* incidence rate is quantitatively small and statistically insignificant for both 3-second play rates and CTRs, although it is positive for 9-second plays. Overall, our results strongly suggesting that the observed heterogeneity in incidence rates is mainly driven by county characteristics, e.g. demo-

7. One concern here may be that the baseline CWA adoption rate differs by regions, which may also drive differential responses to our treatment ads. To alleviate these concerns, we ran a fourth treatment arm simultaneously to our interventions, which included a poll on whether the viewer had already installed the CWA or not. Appendix Figure 3.A.2 shows that, if anything, the relationship between incidence rates and the share of "yes" respondents is negative.

8. The coefficients of all variables included in the LASSO regression are presented in Appendix tables 3.B.10.

graphics and political attitudes, that determine both a county's incidence rates and residents' response to interventions promoting the CWA.

3.3.2 Results from a representative online survey

To complement our results from the social media experiment, we conduct an online survey with a representative sample in Germany (N = 5,830; age of respondents ranging from 18 to 65) between 11 Nov and 17 Dec 2020. 37% of respondents report that the CWA was currently installed on their smartphone, which is roughly consistent with the population-level download statistics conditional on the age groups represented in our sample (Blom, Wenz, Cornesse, et al., 2020). We embedded the same interventions as before in the survey, and additionally elicited beliefs and attitudes toward the CWA both before and after respondents were randomly shown one of the video ads (T0, T1, or T2), hence allowing us to study the behavioral responses to our intervention in more detail. A crucial distinction between this survey setting and the social media setting, apart from sample selection, is the environment in which subjects are exposed to our intervention. Unlike in the social media experiment, subject choose to participate in our survey and thus have, at least to some degree, committed to devote their time and attention to respond to our questions and engage with the intervention, although they were unaware of it ex ante. Therefore, the extensive margin effect that we demonstrated in the social media setting is mostly muted in the survey experiment, which allows us to isolate the intensive margin effect, i.e. the behavioral response given that one engages with the treatment video.

We estimate the effect of the different video conditions using the change in attitudes and beliefs about the CWA after watching the respective video. More specifically, we focus on three outcome variables: (1) a factor variable on favorability of attitudes towards the CWA, constructed from 6 questions on subjects' perception of costs and benefits of app adoption; (2) subjects' willingness to install the CWA, or if already installed, to regularly use the CWA; (3) an incentivized (though low-stakes) choice about which share of a bonus payment of 1 Euro they wish to donate to a marketing campaign for the CWA.[9] For more details, see Materials and Methods and Appendix 3.C. As the T1 and T2 videos generally perform at very similar levels, we will pool the two treatment groups for most of our analyses. Furthermore, as the interventions are specifically targeted at individuals who have not adopted the app so far, we estimate all effects separately for the subsamples of CWA non-adopters, who have yet to download the app, and CWA adopters, who had already downloaded it.

Table 3.3.3 shows that the regular promotion video (T0) alone already significantly improves attitudes towards the CWA for both non-adopters and adopters by

9. This incentivization was truthful. We spent the entire donated budget in May and June 2021 to advertise a functionality upgrade of the CWA on Facebook and Instagram.

Table 3.3.3. Difference-in-differences estimates of survey responses

	CWA non-adopters			*CWA adopters*		
	favorability factor	willing to install	donation share	favorability factor	willing to use	donation share
Treated	0.067***	0.174***	0.020***	0.057***	-0.037	0.009
	(0.022)	(0.049)	(0.007)	(0.020)	(0.046)	(0.008)
Treated × below-avg.	-0.041	-0.171**	-0.000	-0.051	-0.051	-0.001
	(0.031)	(0.069)	(0.010)	(0.031)	(0.067)	(0.011)
below-average status	0.027	0.076	-0.001	0.044*	0.117**	-0.012
	(0.024)	(0.055)	(0.008)	(0.026)	(0.054)	(0.009)
Constant	0.062***	0.109***	-0.010*	0.068***	-0.058	0.003
	(0.017)	(0.038)	(0.005)	(0.017)	(0.039)	(0.007)
Baseline mean	-.393	1.945	0.322	.665	6.509	0.520
Observations	3580	3580	3570	2104	2104	2103
R^2	0.003	0.003	0.004	0.002	0.005	0.004

Standard errors in parentheses are clustered at the county level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6-7% of a standard deviation; it also increases non-adopters' average self-reported willingness to install the app by 0.1 points on a 7-point Likert scale. We acknowledge that part of this may be due to experimenter demand effects, but this is precisely the reason why we use T0 as control group. Importantly, we observe small but statistically significant ($p < 0.01$) treatment effects for non-adopters on all three outcome measures in response to the interventions videos with feedback on local incidence rates. Consistent with the results from the social media experiment, the effects tend to be stronger for individuals who receive unfavorable comparison feedback with regard to the incidence rates in their county. Compared to those who have been assigned the T0 video, non-adopters in above-average incidence counties who have been randomly assigned the T1 or T2 video improve their attitudes towards the CWA by about 4.5% SDs ($p = 0.003$) and increase their donation share by an average of $2\%p$ ($p < 0.001$). They also report a higher willingness to install the app (0.08 points on a 7-point Likert scale, $p = 0.024$). In contrast, non-adopters from below-average incidence counties do not become more willing to install the app after the intervention. For individuals who have already downloaded the CWA, we generally do not observe further positive effects after showing the interventions, with the exception of an increase in the favorability factor by 0.06 SDs ($p < 0.01$) in above-average counties. Note, however, that the baseline attitudes and donations of adopters before watching the videos were already at far higher levels than for non-adopters, which is unsurprising given endogenous selection into app adoption.

Generally, our survey results suggest that combining the promotion video with locally targeted feedback on regional Covid-19 incidence rates actually has a positive effect on their view towards the CWA beyond simply drawing more initial attention, for example by putting consumers in a more open state of mind. However, the point

estimates are relatively small, suggesting that they may only sufficiently convince a subset of individuals who are at the margin of app adoption. Thus, the social media and the survey experiment are remarkably consistent in their conclusion. In both settings, our interventions produce measurably more success in promoting the CWA than the just regular advertisement by itself. In both settings, the overall prospects are dampened by the fact that all effects are quantitatively small on an absolute scale.

It is, of course, possible that the behavioral intervention we test is simply too minimal to produce any quantitatively more satisfying effects, but when we draw on the survey responses to shed more light on this apparent disinterest in our appeals for using the CWA, an even more banal explanation takes the forefront: that the disinterest is in the app itself. In fact, the general willingness to install the app among the pool of non-adopters is so staggeringly low that only about 10% stated that they were at least somewhat likely to install the CWA in the near future, whereas 65% reported that they were completely certain they would not do so (see Appendix Figure 3.A.3). Given that almost two-thirds of our target population seems categorically opposed to what we are trying to sell them, it is unsurprising that we do not find larger responses to our interventions.

3.4 Discussion

More than a year after the start of the Covid-19 pandemic, public compliance with public health measures is faltering in many countries. This also applies for DCT apps, which were rolled out on large scale in the spring and summer of 2020, but whose adoption rates have all but stagnated at insufficient levels by now. Having reached a point where it becomes harder and harder to motivate more people to use DCT app, conventional promotion strategies may have reached a certain saturation point.

In a large-scale social media experiment, we show that a simple behavioral intervention that combines a regular promotion ad for the German CWA with locally targeted feedback on severity of the pandemic can be significantly more effective in increasing consumers' willingness to consider taking-up the app, as indicated by 30% higher link click-through rates compared to the control condition that only shows the promotion video. This is in large parts driven by the extensive margin effect that more users stop and engage with the intervention, as indicated by a 30% higher rate of 3-second plays for the treatment videos relative to the control video. This does not imply that there is no hidden effect on the intensive margin, since the feedback on regional incidence rates may have drawn in a different selection of viewers that are on average less interested in installing the app per se, be it because they already use it or because they do not want to use it. Indeed, in a complementary online survey experiment, we do observe that watching the treatment videos lead to slightly more positive attitudes towards the CWA. But whether or not consumers start paying at-

tention to an intervention in the first place can be of overwhelming importance for its effectiveness in natural everyday context, something that may remain undetected in the more controlled settings of e.g. a lab experiment or an online survey, where subjects have chosen to temporarily remove themselves from other purposes and distractions. For example, information disclosure policies such as for the nutritional value of food items or the energy efficiency of electronic devices may benefit more from salient and colorful labels that draw consumers' attention. The importance of this "stopping effect" has also been emphasized by research in marketing (Drèze and Hussherr, 2003; Pieters and Wedel, 2004; Pieters, Wedel, and Bartra, 2010) and it may hold especially true in this modern digital era, where consumers are constantly exposed to an overabundance of information and stimuli competing for attention.

A somewhat disheartening finding is that the effectiveness of out interventions follow an adverse distributional pattern in the sense that social media users from high-incidence counties tend to respond less to the locally targeted feedback, even though fast and comprehensive contact tracing is particularly important in these counties. This is surprising on the one hand, given that one might expect that learning about the high-incidence rate in one's county would serve as motivational boost to engage in more strongly in public health measures such as CWA use. On the other hand, this is less surprising when taking into consideration potential endogeneity, for example if counties in which residents are more skeptical about public health measures such as the CWA will also tend to exhibit higher Covid-19 incidence rates (see e.g. Fang, Freyer, Ho, et al., 2021). Indeed, we find that this negative pattern is driven by the predictable component of a county's incidence rate, which speaks against behavioral mechanisms like information avoidance. Encouragingly, we also do observe that highlighting that one's county is faring worse than other comparable counties can increase the willingness to seek out more information about the CWA. This heightened response of users who receive disadvantageous feedback is consistent with previous studies on group comparison (Böhm and Rockenbach, 2013; Cárdenas and Mantilla, 2015), as well as with neuropsychological evidence that attention responds more strongly to negative as opposed to positive stimuli (Ito, Larsen, Smith, et al., 1998; Smith, Cacioppo, Larsen, et al., 2003).

Our study has several limitations, the most important one being that we cannot observe actual take-up of the app, and neither do we know the number of downloads by different demographic groups or in different regions. Hence, an experimental design that randomizes promotion messages by region and subsequently compares download statistics, although compelling, would not have been feasible. Our performance measure of link clicks that lead to the CWA website are, at the minimum, a strong indicator for being open to seeking out more information on the CWA, and at best captures the intention to immediately download the app. Thus, we believe using this as main outcome is not without merit. However, we are aware that link clicks might misrepresent the effect on actual uptake of the CWA, although the direction is not clear. On the one hand, users may follow the link but never actually

install the app, which would lead to an overestimation. On the other hand, users may directly look for the CWA app on their app store without following the link first, and there might be longer-term effects of being exposed to the ad that are not captured by immediate link clicks; for example, consumers may be easier persuaded to install the app at the next opportunity. One necessary condition for such persisting effects is that the user has to notice and engage with the ad to at least some degree. As we find strong treatment effects on initial engagement (3-second video plays) and retention (e.g. 9-second video plays) in the social media sample, as well as positive effects on CWA views and uptake intentions in the survey sample, we are confident that our intervention based on locally targeted feedback can improve upon the effectiveness of regular promotion messages in convincing people on the margin to use the app. In a related study on encouraging CWA use that in fact allows tracking of smartphone apps used by participants in an online survey, Munzert, Selb, Gohdes, et al. (2021) find that a small 1 Euro incentive alone can dramatically increase CWA adoption several weeks later, suggesting that some non-adopters are indeed relatively indifferent about the CWA and can thus be easily convinced to use it. In light of this evidence, it seems plausible that our behavioral intervention can also boost actual CWA adoption.

Notwithstanding the imperfections of our outcome measures, one conclusion from our study seems to be that, for all the improvements that our behavioral intervention achieves compared to the regular promotion strategy, the baseline interest in the app remains exceedingly low, with around 1 out of 1000 users following a link to the CWA website, which is not unheard of for social media ads but still performs below average. The representative online survey helps us provide more context for this lack of clout. It shows that, half a year after the CWA was introduced, attitudes in the population towards it have become polarized and entrenched. While a (substantial) minority of the population has installed the app on their smartphone by now — be it happily or grudgingly —, most of the remaining non-adopters exhibit a strong unwillingness to install the app. This leaves only a relatively small share of the population, perhaps 10 to 20% in Germany, who do not have the app yet but can still be somewhat easily convinced to start using it. Behavioral interventions (or nudges) are, by definition, designed in a way that it influences people who are either relatively indifferent to a cause or who are in fact inclined to follow the cause but had not done so yet due to some behavioral barrier like inertia or self-control problems. Thus, any such promotion strategy aimed at encouraging adoption of DCT apps will unfold its effect mostly on these people at the margin, and the thinner this margin becomes, the less effective the intervention will become overall. As the most amenable people all start using the app over time, further increasing adoption rates will become more and more challenging, since the remaining pool of non-adopters

is increasingly composed of the most skeptical and reluctant people, who have very low perceived benefit-cost ratios of using the app.[10]

While our behavioral intervention does unfold some positive effects on willingness to use DCT apps on the margin, the prospects of reaching adoption rates close to ex ante targeted levels seem bleak, with results from our representative survey suggesting that is it highly unlikely that more than 40% — not to speak of 60% — of the population in Germany will eventually use the CWA. Similar problems are expected in other countries. Our empirical findings are also consistent with Munzert et al. (2021), who (among others) test the effectiveness of financial incentives in motivating CWA adoption. Their intervention was conducted in August 2020, a few months prior to ours, and the baseline adoption rate thus lower at that time (about 18 million downloads), but their results also point to a growing entrenchment in attitudes. Munzert et al. (2021) find even a modest 1 Euro monetary incentive increase tracked app uptake by about 18%p among non-adopters four weeks later, but that this effect is virtually the same for a 5 Euro incentive (20%p) — lending support to the interpretation that there is a group of people on the margin who are relatively indifferent and thus easily convinced to use the app, but that by now the majority of non-adopters is very reluctant and unlikely to be swayed. This also fall in line with an early study by Blom et al. (2020), who projected an adoption rate of 37.9 percent of the population aged 18-77 in Germany based on survey responses from June 2020.

Our survey results further show that non-adopters on the margin mainly differ from CWA adopters in their concerns about data protection. While the overall prevalence of privacy concerns is surprising given all efforts to make everything as transparent and secure as possible, it does imply that loosening data protection standards in order to enhance the app's functionality, e.g. by forwarding data to health agencies, would likely backfire by eroding trust both among adopters and non-adopters, and should only be implemented as optional user choice. The relative success of our behavioral intervention also suggests that another route for DCT apps could be to add ancillary benefits, for example by serving as information hub for regional Covid-19 infection situations and current public health regulations. Finally, studies on the acceptance of DCT apps can also inform strategies to promote vaccination in the population, as there many parallels. Both DCT apps and vaccination become more effective the larger the share of adopters in the population; both require individuals' trust in government, public health institutions, and app developers as well as medical researchers; both may induce fear of hidden costs, e.g. privacy concerns and fear of surveillance in case of CWA, and hidden (long-term) side effects in the case of vaccines; lack of perceived benefits among groups who are at low health risk. Indeed, our survey data reveals an especially strong relationship between CWA

10. This is reminiscent of classical frameworks of technology diffusion (Rogers, 1962).

use and attitudes toward vaccination above and beyond general attitudes towards public health measures.[11] One differences may be that for vaccines, uncertainty on side effects can unravel over time, when more people get vaccinated, there is also potential for misinformation based on happenstance events such as patients dying shortly after vaccination, even if for completely unrelated reasons. While some early surveys indicated widespread support for use of DCT apps (Altmann, Milsom, Zillessen, et al., 2020), the adoption rates that were hoped for have not been reached until now, providing a cautionary tale for vaccination efforts across the globe and stressing the importance of formulating strategies to overcome vaccine hesitancy.

3.5 Materials and methods

Social media experiment

Study Design. In our Facebook study, we made 1,368,709 impressions (number of times an instance of an ad is on screen for the first time) in Germany using Facebook Ads (with an estimated reach of 1,115,404 unique users) between October 07, 2020, and October 17, 2020. We used Facebook's A/B testing feature to randomly assign impressions to the different intervention groups. T0 is our control group, which is shown the conventional CWA promotional video which does not contain any county specific information. For the treatment interventions (T1 and T2), we provided information about the 7-day incidence rate, whether the incidence rate is above/below the state average and by how many percent, for the German county that the Facebook user resided in. T2 is shown an additional message - "make sure this changes" if the county has incidence rates above the state average and "make sure it stays this way" if the county has incidence rates below the state average. County specific information (incidence rates and state comparisons) in the treatment interventions was updated 3 other times on October 09, 12, and 14, 2020. Targeting of the county level interventions was done by matching zip codes contained within each of the 221 counties in our sample. Other than geographical targeting, we do not target the advertisements using any other criteria. We also used reach as the campaign objective for our advertisements, which allows Facebook to show the advertisements to as many Facebook users as possible.

We downloaded advertisement performance data that included the number of link clicks, 3 second video views and impressions (aggregated at the county-day-treatment level) from Facebook. The dataset was then disaggregated such that the unit of observation was an impression (i.e., whether a single impression led to a link click or a 3 second video play). However, the joint distribution is unknown to us (i.e., we do not know whether the same impression led to a link click and a 3 second video play), so we did not make use of any joint distribution in the analysis.

Analysis. We use a linear probability model to estimate the average treatment effect of our feedback intervention by using the following regression equation:

11. The correlation coefficient between CWA adoption and vaccination attitudes is $\rho = 0.3099$, far stronger than the correlation of CWA use with any other public health measures (correlations range from 0.1062 to 0.2138). See Appendix Figure 3.A.4.

$$Y_{i,c} = \beta_0 + \beta_1 Treat_{i,c} + \gamma_c + \epsilon_{i,c}$$

where $Y_{i,c,t}$ is the outcome variable — 3 second video play, 9 second video play, or link click — for impression i in county c, $Treat_{i,c,t}$ is an indicator whether the impression is in one of the treatment groups (T1 or T2), γ_c represents county fixed effects, and $\epsilon_{i,c}$ is the error term. We use heteroscedastic robust SEs, clustering by counties. Coefficient β_1 can be interpreted as the causal effect of the treatment because we have randomization across treatment groups. To estimate the heterogeneity of the treatment effect with respect to 7-day incidence rates and whether the county's incidence rate is above/below the state average (shown in the treatment interventions), we use the following regression equation:

$$\begin{aligned} Y_{i,c,t} = \beta_0 + \beta_1 Treat_{i,c,t} + \beta_2 AboveStateAvg_{c,t} + \beta_3 IncidenceRate_{c,t} + \\ \beta_4 AboveStateAvg_{c,t} \times Treat_{i,c,t} + \beta_4 IncidenceRate_{c,t} \times Treat_{i,c,t} + \gamma_c + \delta_t + \epsilon_{i,c,t} \end{aligned}$$

where $AboveStateAvg_{c,t}$ is an indicator for whether impression i came from a county c on day t that has 7-day incidence rates above the state average and $IncidenceRate_{c,t}$ is the 7-day incidence rate for county c that impression i belongs to, on day t.

Predicted Incidence Rates. To decompose the 7-day incidence rates into observable county characteristics and unobservable characteristics/random variation, we included a host of county level characteristics (demographics, voting shares, GDP, etc.) to predict incidence rates using LASSO regression with 10-fold cross validation. We then estimate the following regression specification:

$$\begin{aligned} Y_{i,c,t} = \beta_0 + \beta_1 Treat_{i,c,t} + \beta_2 AboveStateAvg_{c,t} + \beta_3 IncidenceRateResidual_{c,t} \\ \beta_4 AboveStateAvg_{c,t} \times Treat_{i,c,t} + \beta_5 \widehat{IncidenceRate}_c \times Treat_{i,c,t} + \\ \beta_6 IncidenceRateResidual_{c,t} \times Treat_{i,c,t} + \gamma_c + \delta_t + \epsilon_{i,c,t} \end{aligned}$$

where $\widehat{IncidenceRate}_c$ is the predicted 7-day incidence rates from the LASSO for county c that impression i belongs to. $IncidenceRateResidual_{c,t}$ is the difference between the actual incidence rates and the predicted incidence rates. Variable selection methods such as the LASSO can be unstable and have considerable model uncertainty (i.e., small changes in the dataset can yield very different models after selection) and could produce fitted values that have large variance. Hence, we use bootstrapping to reliably estimate the standard deviation of the coefficients in this regression specification that contains predicted incidence rates as explanatory variables. We first sample all 401 German counties with replacement and predict the incidence rates using the full set of county characteristics. We then estimate this regression equation using only the unique counties from the bootstrapped sample and their respective resampling weights, and we repeat this bootstrap procedure for 10,000 iterations.

Online survey

Study Design. In the online survey, 6,000 German participants were recruited through the online market research firm, Dynata from 11 November to 17 December 2020. Participants were sampled using pre-specified distributions based on age, gender, and county, and were randomly assigned to one of the 3 groups, T0, T1, and T2, which are shown the same video advertisements as in the Facebook study. The county specific information (7-day incidence rates and state comparisons) shown in the treatment interventions were updated daily. We

include 5,830 responses that met our criteria for inclusion in the analysis, namely no speeding, failing attention check questions, and giving straight line responses. Compensation to participants included a fixed component determined by the survey company and a variable component of 1 Euro. Participants had a 50% chance of being selected for the variable component and if selected, they could decide how much of an additional 1 Euro they wish to keep and how much to donate towards marketing efforts for the German CWA app.

Participants were first asked questions about their demographics, personality and preferences, regional identification, political attitudes, public health attitudes, attitudes towards Covid, and media/news consumption. They were then shown a video depending on which group they were in – T0 were shown the conventional CWA ad, participants in the T1 and T2 were provided with additional county-specific information in the videos, and T2 was further shown an additional message depending on whether the county has incidence rates above the state average (similar to the Facebook study). Lastly, they were asked about their attitudes on the CWA app and willingness to install or use the app (7-point Likert scale) both before and after the video intervention.

Analysis (Online Survey). We estimate the following equation:

$$\triangle Y_{ic} = \alpha + \beta_1 Treat_i + \gamma_0 Below_{ic} + \gamma_1 Treat_i \times Below_{ic} + \epsilon_{ic}$$

where $\triangle Y_i$ is the difference between the outcome variable before and after the video intervention for individual i. Outcomes that we focus on include favorability of attitudes towards the CWA app (normalized to mean 0 and standard deviation 1), willingness to install/use the CWA app (on 7-point Likert scale), and the share of \$1 that individual i is willing to donate to aid marketing efforts for the CWA app on Facebook. $Treat_i$ is an indicator for if individual i is in one of the treatment groups (T1 or T2). $Below_{ic}$ is an indicator for whether the county of residence c had a below-average incidence rate relative to other counties in the state.

Appendix 3.A Supplementary figures

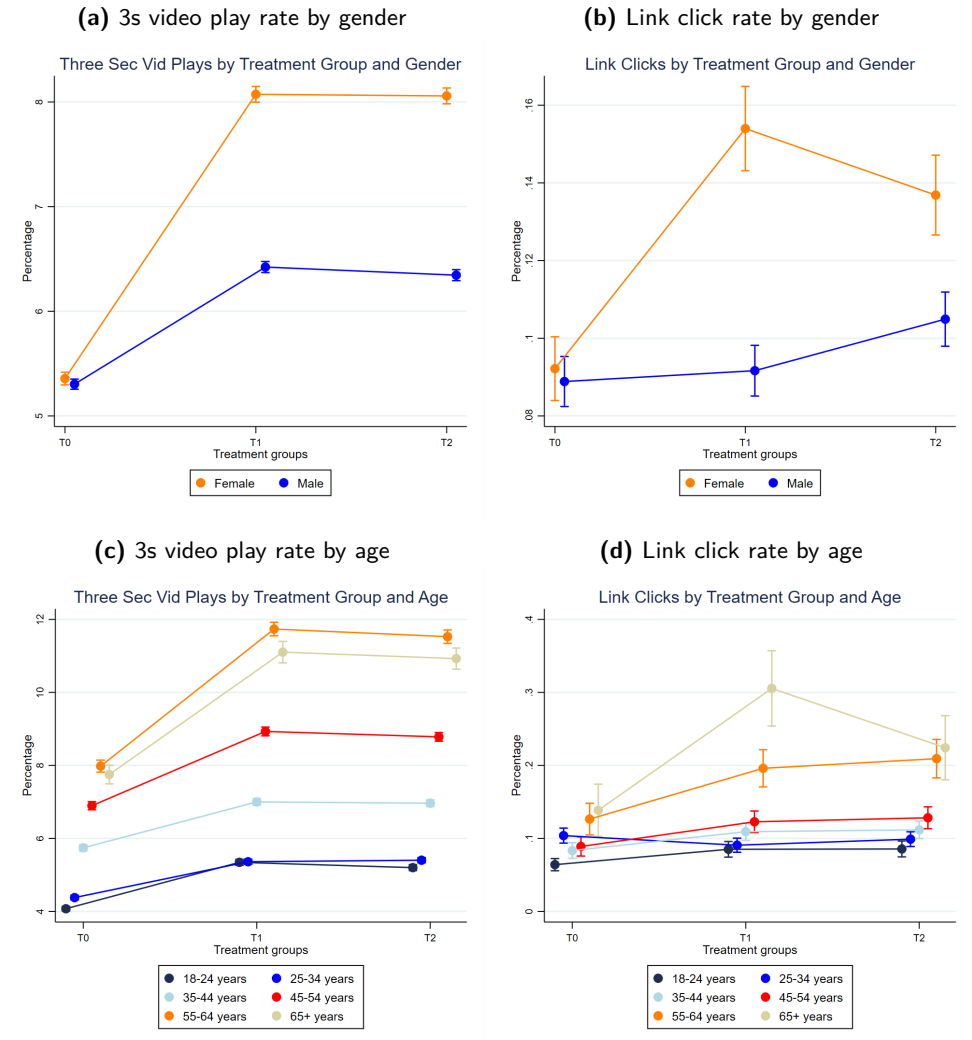


Figure 3.A.1. Heterogeneity in ad performance by age and gender

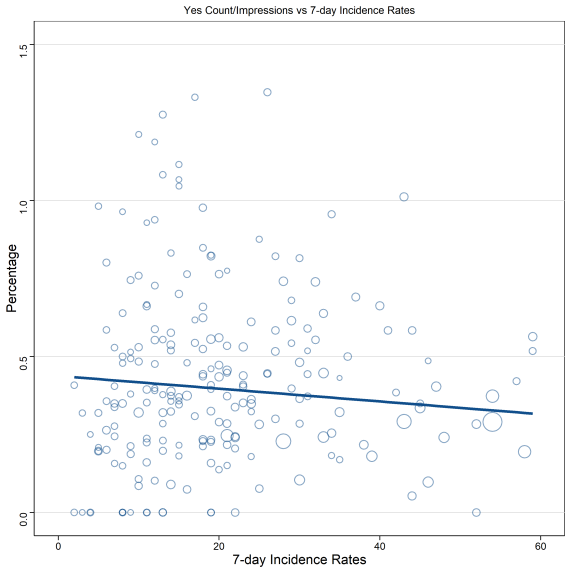


Figure 3.A.2. Share of CWA adoption responses per impression in Poll condition

Notes: The figure plots the county-level percentage of uses who responded with "yes" in the CWA poll (relative to the overall number of impressions) by 7-day incidence rate in the county.

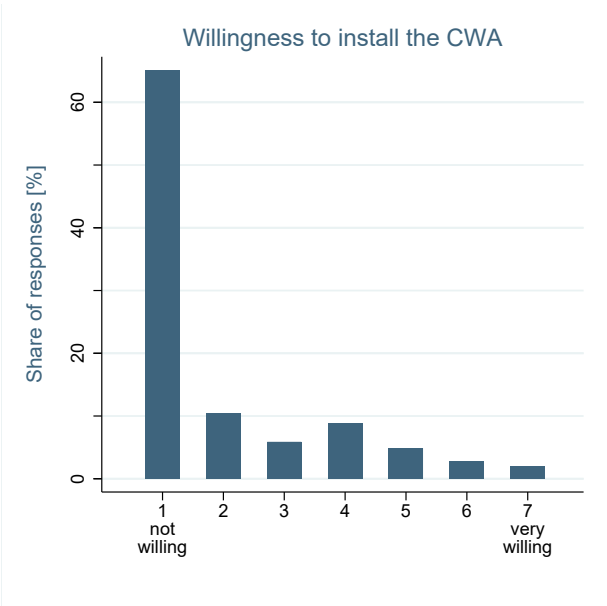


Figure 3.A.3. Histogram of baseline willingness to install the CWA among non-adopters

Figure 3.A.4. Correlations between CWA attitudes and other public health attitudes

Appendix 3.B Supplementary tables

Table 3.B.1. Average Treatment Effects

	3s plays [%]	9s plays [%]	CTR [%]	$\frac{\text{CTR}}{\text{3s plays}}$ [%]
T1 video	1.717***	1.811***	0.025***	-0.081
	(0.127)	(0.057)	(0.008)	(0.124)
T2 video	1.660***	1.108***	0.028***	-0.009
	(0.135)	(0.081)	(0.008)	(0.109)
Constant	5.325***	1.050***	0.090***	1.688***
	(0.084)	(0.041)	(0.004)	(0.066)
County fixed effects	*yes*	*yes*	*yes*	*yes*
$H_0 : T1 = T2$	$p = 0.396$	$p < 0.001$	$p = 0.728$	$p = 0.560$
Impressions	1044271	1044271	1044271	1044271
Counties	221	221	221	221
R^2	0.790	0.811	0.395	0.344

Standard errors in parentheses (clustered at county level). $^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Table 3.B.2. Heterogeneity in Incidence Rates (Split Treatments)

Panel A: 7-day Incidence Rates			
	3s Video Plays	9s Video Plays	Link Clicks
T1	2.2685***	1.8560***	0.0443***
	(0.1465)	(0.0827)	(0.0140)
T2	2.1789***	1.4490***	0.04313***
	(0.1575)	(0.0891)	(0.0139)
Above State Average	-0.0899	0.2079**	0.0141***
	(0.2213)	(0.0982)	(0.0225)
7-day Incidence (Cases per 1000)	2.8680***	0.7471*	0.1200*
	(0.6606)	(0.3975)	(0.0695)
T1 × Above State Average	1.0642***	-0.3539**	0.0367*
	(0.2556)	(0.1394)	(0.0201)
T2 × Above State Average	1.0862***	0.4789***	0.0350*
	(0.2754)	(0.1602)	(0.0198)
T1 × 7-day Incidence (Cases per 1000)	-3.9243***	0.5193	-0.1351***
	(0.6004)	(0.3363)	(0.0507)
T2 × 7-day Incidence (Cases per 1000)	-3.8546***	-2.059***	-0.1175**
	(0.6813)	(0.3916)	(0.0525)
Panel B: Predicted 7-day Incidence Rates			
	3s Video Plays	9s Video Plays	Link Clicks
T1	3.0289***	2.3501***	0.0615***
	(0.1962)	(0.1139)	(0.0183)
T2	3.1221***	1.9380***	0.0768***
	(0.2173)	(0.1190)	(0.0185)
Above State Average	-0.1972	0.1410*	0.0117
	(0.2187)	(0.0937)	(0.0226)
7-day Incidence (Residual)	0.5291	-0.5969**	0.0426
	(0.5391)	(0.3319)	(0.0578)
T1 × Above State Average	1.1798***	-0.2436**	0.0389**
	(0.2480)	(0.1197)	(0.0205)
T2 × Above State Average	1.2656***	0.5643***	0.0408**
	(0.2606)	(0.1507)	(0.0204)
T1 × 7-day Incidence (Predicted)	-6.4150***	-1.1541***	-0.1908***
	(0.7511)	(0.4228)	(0.0662)
T2 × 7-day Incidence (Predicted)	-7.0181***	-3.6825***	-0.2289***
	(0.8409)	(0.4579)	(0.0669)
T1 × 7-day Incidence (Residual)	-0.4612	2.5665***	-0.0543
	(0.7808)	(0.3934)	(0.0595)
T2 × 7-day Incidence (Residual)	0.1439	0.0769	0.0314
	(0.6647)	(0.4019)	(0.0624)
Unconditional Mean for T0	5.3307	1.0475	0.0897
Observations	1044271	1044271	1044271
County Fixed Effects	Yes	Yes	Yes
Day Fixed Effects	Yes	Yes	Yes

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
Standard errors in Panel A are clustered at the county level
Standard errors in Panel B are obtained from bootstrapping (10000 iterations)

Table 3.B.4. Intensive Margin Effects ($\frac{link\ clicks}{3s\ video\ plays}$)

	(1)	(2)	(3)
T1	-0.0825	0.2338	0.1405
	(0.1253)	(0.2284)	(0.2557)
T2	-0.0269	0.2790	0.6613**
	(0.1079)	(0.2212)	(0.2716)
7-day Incidence (Cases per 1000)		1.0541**	
		(0.4826)	
7-day Incidence (Predicted)			0.1709
			(0.7039)
7-day Incidence (Residual)			0.5147
			(1.2432)
T1 × 7-day Incidence (Cases per 1000)		-1.0909*	
		(0.6408)	
T2 × 7-day Incidence (Cases per 1000)		-1.0548	
		(0.6401)	
T1 × 7-day Incidence (Predicted)			-0.0580
			(0.9399)
T2 × 7-day Incidence (Predicted)			-1.1925
			(0.0.9923)
T1 × 7-day Incidence (Residual)			-0.9486
			(1.5316))
T2 × 7-day Incidence (Residual)			1.6612
			(1.5548)
Constant	1.6935***	1.3878***	1.5218***
	(0.0908)	(0.1675)	(0.1907)
Observations	663	663	663

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.6. Clicks (all)

	(1)	(2)	(3)
Treatment	0.4379***	0.6732***	0.9139***
	(0.0381)	(0.0413)	(0.0592)
Above State Avg		0.0204	-0.0028
		(0.0546)	(0.0539)
7-day Incidence (Cases per 1000)		1.0565**	
		(0.1830)	
7-day Incidence (Residual)			0.3232**
			(0.1465)
Treatment × Above State Avg		0.2094***	0.2521***
		(0.0711)	(0.0660)
Treatment × 7-day Incidence (Cases per 1000)		-1.2040***	
		(0.1612)	
Treatment × 7-day Incidence (Predicted)			-2.0061***
			(0.2046)
Treatment × 7-day Incidence (Residual)			-0.1607
			(0.1620))
Constant	0.1604***	-0.1564***	0.1707***
	(0.0253)	(0.0511)	(0.0312)
Observations	1044271	1044271	1044271
County Fixed Effects	Yes	Yes	Yes
Day Fixed Effects	Yes	Yes	Yes

Standard errors in parentheses and clustered at the county level
Standard errors in column 3 are obtained from 1000 iteration bootstrap
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.8. Clicks (all) Intensive Margin Effects ($\frac{clicks\ all}{3s\ video\ plays} \times 100$)

	(1)	(2)	(3)
Treatment	5.0184***	8.2124***	11.0265***
	(0.3938)	(0.4797)	(0.5685)
7-day Incidence (Actual)		3.1269**	
		(0.7695)	
7-day Incidence (Predicted)		3.5368**	
		(0.9684)	
7-day Incidence (Residual)			0.9743
			(1.7106)
Treatment × 7-day Incidence (Actual)		-11.0713***	
		(1.2842)	
Treatment × 7-day Incidence (Predicted)			-18.2921***
			(2.0397)
Treatment × 7-day Incidence (Residual)			0.2731
			(3.0635))
Constant	3.0699***	2.1631***	1.8811***
	(0.1512)	(0.2365)	(0.2509)
Observations	663	663	663

Standard errors in parentheses and clustered at the county level
Standard errors in column 3 are obtained from 1000 iteration bootstrap
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.B.10. Lasso Variables

Variable	Coefficient
Intercept	19.2374
Area	-0.4968
Log population	0.8177
Population density	4.8878
Share of 15-24 year olds	3.0836
Share of 25-29 year olds	-1.2869
Share of 40-44 year olds	0.0744
Share of 45-49 year olds	0.5940
Share of female 0-14 year olds	0.4651
Share of female 35-39 year olds	-2.7644
Share of female 35-39 year olds	2.9143
Share of female 50-54 year olds	-0.7190
Share of female 65-74 year olds	-0.0427
Share of females	0.7823
Share of non-germans	8.1287
2017 election turnout	0.6438
2017 share voted for cdu	2.3881
2017 share voted for green	-2.6689
2017 share voted for linke	2.5230
2019 election turnout	0.8787
2019 share voted for fdp	0.7803
2019 share voted for linke	0.3957
2019 share voted for afd	0.7407
Unemployed	-3.2495
Unemployed - longterm	5.0242
Unemployment Rate (female)	-1.2521
Unemployment Rate (non-German)	1.3785
Unemployment Rate (youths)	-1.3904
Unemployment Rate (long-term)	1.3534
Education (secondary schools)	2.0485
Education (secondary school)	2.0794
Education (high school)	2.6705
Employed	0.7973
Employed (farming)	0.9380
Employed (industry)	2.2837
Employed (building)	0.5623
Employed (financial)	-0.7856
Log GDP per worker	-1.3199
GDP (building)	1.4752
GDP (financial)	0.7736
Income per capita	-0.2999
Hospitals	-0.8072
Hospital beds	0.2730
Urban	-0.7316
East German	0.0600

Adaptive Lasso estimates using first stage OLS regression coefficients as weights. Right hand side variables are standardized.

Appendix 3.C Survey measures

In this section, we list the survey questions used to elicit subjects' preferences and pandemic-related behavior. All questions used to elicit preferences prompted subjects to indicate their self-assessment on a ten-point Likert scale. We indicate alongside each question the corresponding preference (not displayed in the survey). Note that there are two questions regarding direct negative reciprocity; we give each of them equal weight to obtain a single measure for this preference. The questions used to elicit pandemic-related behavior prompted subjects to indicate their self-assessment on a seven-point Likert scale.

3.C.1 Corona Warn App attitudes

Is the Corona Warn App currently installed on your mobile phone?
Yes
No

How much do you agree with the following statements about the Corona Warn App? Please answer on a scale from 1 to 7. A value of 1 means: Strongly disagree. The value 7 means: I fully agree.

"The Corona Warn App ...

... helps to slow down the spread of Covid-19 in Germany."

... helps to slow down the spread of Covid-19 in my county."

... is of little use to me personally."

... is questionable in terms of data protection."

... is a good tool for tracking infection chains."

... is not used by enough people yet."

If CWA not installed yet:
How likely is it that you will install the Corona Warn App on your mobile within the next 7 days?
Please answer on a scale from 1 to 7. The value 1 means: very unlikely. The value 7 means: very likely.

If CWA already installed:
How likely is it that you will report a test result to the Corona Warn App in case it is positive?
Please answer on a scale from 1 to 7. The value 1 means: very unlikely. The value 7 means: very likely.

3.C.2 Public health behavior

To what extent do the following statements apply to your own behavior? Please rate again on a scale from 1 to 7. The value 1 means: does not apply at all. The value 7 means: applies completely.

I keep a distance of at least 1.5 meters from other people.

I will isolate myself socially if I have had contact with an infected person.

I always keep myself up to date with news about the corona pandemic.

I wash and disinfect my hands regularly.

I will get vaccinated against the coronavirus when a vaccine becomes available.

I cough and sneeze into the crook of my elbow.

I wear mouth and nose protection in public.

I ventilate regularly when several people are using a room.

I avoid social contacts as much as possible.

I will inform other people if I have been infected with the coronavirus.

References

Adstage. 2020. "Paid Media Benchmark Report Q1 2020." [132]

Ahmed, Nadeem, Regio A. Michelin, Wanli Xue, Sushmita Ruj, Robert Malaney, Salil S. Kanhere, Aruna Seneviratne, Wen Hu, Helge Janicke, and Sanjay K. Jha. 2020. "A Survey of COVID-19 Contact Tracing Apps." *IEEE Access* 8: 134577–601. [125]

Altmann, Samuel, Luke Milsom, Hannah Zillessen, Raffele Baslone, Frederic Gerdon, Ruben Bach, Frauke Kreuter, Nosenzo, Daniele, Toussaert, Severine, and Johannes Abeler. 2020. "Acceptability of App-Based Contact Tracing for COVID-19: Cross-Country Survey Study." *JMIR mHealth and uHealth* 8 (8): [143]

Blom, Annelies G., Alexander Wenz, Carina Cornesse, Tobias Rettig, Marina Fikel, Sabine Friedel, Sebastian Juhl, Roni Lehrer, Katja Möhring, Elias Naumann, Maximiliane Reifenscheid, and Ulrich Krieger. 2020. "Barriers to the Large-Scale Adoption of the COVID-19 Contact-Tracing App in Germany." *Working Paper*, [137, 142]

Böhm, Robert, and Bettina Rockenbach. 2013. "The inter-group comparison-intra-group cooperation hypothesis: comparisons between groups increase efficiency in public goods provision." *PloS one* 8 (2): e56152. [127, 140]

Cárdenas, Juan C., and César Mantilla. 2015. "Between-group competition, intra-group cooperation and relative performance." *Frontiers in behavioral neuroscience* 9: 33. [140]

Charness, G., L. Rigotti, and A. Rustichini. 2007. "Individual behavior and group membership." *American Economic Review* 97 (4): 1340–1352. [127]

Charness, Gary, and Patrick Holder. 2019. "Charity in the Laboratory: Matching, Competition, and Group Identity." *Management Science* 65 (3): 1398–1407. [127]

Chen, Yan, and Sherry Xin Li. 2009. "Group identity and social preferences." *American Economic Review* 99 (1): 431–457. [127]

Contreras, Sebastian, Jonas Dehning, Matthias Loidolt, Johannes Zierenberg, F. Paul Spitzner, Jorge H. Urrea-Quintero, Sebastian B. Mohr, Michael Wilczek, Michael Wibral, and Viola Priesemann. 2021. "The challenges of containing SARS-CoV-2 via test-trace-and-isolate." *Nature communications* 12 (1): 378. [125]

Drèze, Xavier, and François-Xavier Hussherr. 2003. "Internet advertising: Is anybody watching?" *Journal of Interactive Marketing* 17 (4): 8–23. [128, 140]

Facebook. 2015. "The Value of Video for Brands." [132]

Fang, Ximeng, Timo Freyer, Chui Yee Ho, Zihua Chen, and Lorenz Goette. 2021. "Prosociality predicts individual behavior and collective outcomes in the COVID-19 pandemic." *CRC TR 224 Discussion Paper No. 319*, [140]

Ferraro, Paul J., and Michael K. Price. 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95 (1): 64–73. [127]

Ferretti, Luca, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. 2020. "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing." *Science* 368 (6491): [125]

Fetzer, Thiemo, and Thomas Graeber. 2021. "Measuring the scientific effectiveness of contact tracing: Evidence from a natural experiment." *Proceedings of the National Academy of Sciences* 118 (33): e2100814118. [125]

Goette, Lorenz, David Huffman, Stephan Meier, and Matthias Sutter. 2012. "Competition Between Organizational Groups: Its Impact on Altruistic and Antisocial Motivations." *Management Science* 58 (5): 948–960. [127]

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information Avoidance." *Journal of Economic Literature* 55 (1): 96–135. [128, 136]

Golman, Russell, George Loewenstein, Andras Molnar, and Silvia Saccardo. 2020. "The Demand for, and Avoidance of, Information." *Working Paper*, [128, 136]

Ito, Tiffany A., Jeff T. Larsen, N. Kyle Smith, and John T. Cacioppo. 1998. "Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations." *Journal of Personality and Social Psychology* 75 (4): 887–900. [140]

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263–291. [128]

Karlsson, Niklas, George Loewenstein, and Duane Seppi. 2009. "The ostrich effect: Selective attention to information." *Journal of Risk and Uncertainty* 38 (2): 95–115. [136]

Kessler, Judd B., and Katherine L. Milkman. 2018. "Identity in Charitable Giving." *Management Science* 64 (2): 845–859. [128]

Kretzschmar, Mirjam E., Ganna Rozhnova, Martin C. J. Bootsma, Michiel van Boven, Janneke H. H. M. van de Wijgert, and Marc J. M. Bonten. 2020. "Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study." *The Lancet Public Health* 5 (8): e452–e459. [125]

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355. [133]

Mosoff, Ryan, Tommy Friedlich, Teresa Scassa, Kelly Bronson, and Jason Millar. 2020. "Global Pandemic App Watch (GPAW): COVID-19 Exposure Notification and Contact Tracing Apps." [126]

Munzert, Simon, Peter Selb, Anita Gohdes, Lukas Stoetzer, and Will Lowe. 2021. "Tracking and promoting the usage of a COVID-19 contact tracing app." *Nature human behaviour* 5: 247–255. [141, 142]

O'Neill, P. H., T. Ryan-Mosley, and B. Johnson. 2020. "A flood of coronavirus apps are tracking us. Now it's time to keep track of them." Edited by MIT Technology Review. [125]

Pieters, Rik, and Michel Wedel. 2004. "Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects." *Journal of Marketing* 68: 36–50. [128, 140]

Pieters, Rik, Michel Wedel, and Rajeev Bartra. 2010. "The Stopping Power of Advertising: Measures and Effects of Visual Complexity." *Journal of Marketing* 74: 48–60. [128, 140]

Pollmann, Tina R., Julia Pollmann, Christoph Wiesinger, Christian Haack, Lolian Shtembari, Andrea Turcati, Birgit Neumair, Stephan Meighen-Berger, Giovanni Zattera, Matthias Neumair, Uljana Apel, Augustine Okolie, Johannes Mueller, Stefan Schoenert, and Elisa Resconi. 2020. "The impact of digital contact tracing on the SARS-CoV-2 pandemic - a comprehensive modelling study." [125]

Rogers, Everett. 1962. *Diffusion of Innovations*. New York: Free Press of Glencoe. [142]

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* 18 (5): 429–434. [126, 128]

Small, Deborah A., and George Loewenstein. 2003. "Helping a Victim or Helping the Victim: Altruism and Identifiability." *Journal of Risk and Uncertainty* 26 (1): 5–16. [127]

Smith, Kyle N., John T. Cacioppo, Jeff T. Larsen, and Tanya L. Chartrand. 2003. "May I have your attention, please: Electrocortical responses to positive and negative stimuli." *Neuropsychologia* 41: 171–183. [140]

Texeira, Thales, Michel Wedel, and Rik Pieters. 2012. "Emotion-Induced Engagement in Internet Video Advertisements." *Journal of Marketing Research* 49 (2): 144–159. [128]

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake. 2018. "Overcoming salience bias: how real-time feedback fosters resource conservation." *Management Science* 64 (3): 1458–1476. [127]

Wordstream. 2020. "Facebook Ad Benchmarks for Your Industry [2019]." [132]

Wymant, Chris, Luca Ferretti, Daphne Tsallis, Marcos Charalambides, Lucie Abeler-Dörner, David Bonsall, Robert Hinch, Michelle Kendall, Luke Milsom, Matthew Ayres, Chris Holmes, Mark Briers, and Christophe Fraser. 2021. "The epidemiological impact of the NHS COVID-19 app." *Nature* 594 (7863): 408–412. [125]

Chapter 4

The Effect of Transparency on Performance Evaluation in Committees – Evidence from Professional Figure Skating

Joint with Chui Yee Ho

4.1 Introduction

Many high-stakes decisions are undertaken by groups of people, as opposed to a single individual. This includes, among many other examples, the passage of laws by parliaments, implementation of government policies and regulation, judicial rulings by panels of jurors or judges, hiring decisions in the labor market, and performance evaluation in professional sports. By collecting the views of several individuals, committees can potentially aggregate more information, while simultaneously reducing the influence of idiosyncratic biases or preferences of any single evaluator.

However, the quality of decision-making in committees depends on various institutional features and the strategic incentives they generate. One important feature is whether the votes and opinions of each member are made public or kept secret. On the one hand, larger transparency of the decision-making process allows the public to hold individual committee members accountable, who may in turn attempt to become more impartial and put in more effort to acquire relevant information. On the other hand, transparency can result in less effective information aggregation if it exposes members to undesired outside influences and if it causes excessive conformity or conservatism, e.g. members shying away from expressing controversial opinions

and deviating from a norm or consensus.[1] Committee members' responses to higher transparency could thus be ambiguous and nuanced. Yet, with a few notable exceptions (e.g. Meade and Stasavage, 2008; Benesch, Bütler, and Hofer, 2018; Hansen, McMahon, and Prat, 2018), causal evidence on the effects of transparency on real-world committee behavior remains scarce, mainly due to lack of data and other empirical challenges.

In this paper, we study a transparency reform in professional figure skating that made the performance scores awarded by individual committee members publicly visible. Figure skating is an inherently subjective sport, in which the quality of a performance is partially derived from artistic aspects such as music interpretation and choreography. Hence, skaters performances are evaluated by a panel of (typically 7-9) judges. Prior to the 2016-17 season, judges' scores in most competitions were published anonymously, meaning that the distribution of scores and the identities of individual judges were known, but could not be linked to each other. From the 2016-17 season onwards, each judges' scores were published openly, thus enabling public monitoring of individual judges.[2] We examine the effects of higher transparency on judges' evaluation behavior in a difference-in-differences design, using as control group a subset of events (Junior Grand Prix competitions) in which judges' scores were already published openly pre-reform.

This setting allows us to overcome several empirical challenges. First, we observe a large number of comparable decisions by professional evaluators in a high-stakes context, both under anonymous and under transparent disclosure regimes. Second, we know the aggregation mechanism and observe all individual inputs that contribute to the overall decision. Third, we can rule out information transmission and strategic reporting agreements within the committee, as judges are not allowed to communicate with each other pre-decision. Finally, the difference-in-differences setup allows us to control for general time trends unrelated to the reform, thus helping us to isolate the effect of higher transparency.

Generally, individuals have been found to shift their behavior more towards the socially acceptable norm when (feeling) observed by others.[3] Accordingly, if judges

1. An extreme example of group conformity overruling reason has been provided in the famous experiment by Asch (1951). Similarly, it has been argued that the wisdom-of-crowds phenomenon may not hold when the aggregated judgements are not independent but exposed to social influence (Lorenz, Rauhut, Schweitzer, et al., 2011).

2. The reform was implemented due to allegations of nationalistically biased judging at the 2014 Sochi Olympics Women's competition. Calls for de-anonymized publication were at the 2014 ISU (International Skating Union) Congress, and the majority of members voted in favor of the proposal (30 in favor, 24 against, 2 abstentions). Still, it was not until two years later, at the 2016 ISU Congress, that the reform was passed.

3. For example, students tend to reduce (visible) schooling investments when their rankings are revealed to their classmates (Bursztyn and Jensen, 2015), grocery store workers work harder when observed by more productive co-workers (Mas and Moretti, 2009), individuals are more likely to vote if they believe that their voting status would be revealed to their neighbors (Gerber, Green, and Larimer, 2008).

want to appear competent and impartial in the public eye, then higher transparency could trigger judges' reputation concerns and thereby induce them to provide more accurate evaluations (see e.g. Suurmond, Swank, and Visser, 2004; Bar-Isaac, 2012; Gersbach and Hahn, 2012; Hansen, McMahon, and Prat, 2018; Mattozzi and Nakaguma, 2019; Swank and Visser, 2021). This may be of particular importance due to the presence of subjective biases and favoritism in evaluation decisions, which has been well documented in figure skating and beyond.[4] However, there is no objective metric against which the evaluation decisions can be validated against, so the "true" score is never fully revealed — which is of course the very reason why figure skating performances are evaluated by a panel of expert judges.[5] Thus, a natural benchmark for inputs of individual committee members is the comparison to inputs by the other members. This creates strategic incentives for judges to become more "conformist", i.e. to report scores that are closer to the scores that (they think) other judges will report, potentially at the loss of information value.[6]

To explore the potential effects of transparency in our context, we present a theoretical model based on a beauty contest framework à la Morris and Shin (2002) with endogenous information acquisition. Judges are partially motivated by a truth-telling motive, but they also have a distortion motive due to subjective biases, e.g. nationalistic favoritism. Additionally, reputation-concerned judges have a conformity motive, i.e. they want to award scores that are similar to those of their fellow judges. We interpret higher transparency as an increase in this conformity motive. The model highlights three key mechanisms through which transparency can affect judge evaluation behavior. Firstly, judges may exert higher effort to generate more precise signals, as a reduction in noise will generally lead to higher correlation of signals within the panel. Secondly, judges may become more cautious and conservative in their scores, e.g. by anchoring towards a common prior, thus leading them to place lower weight on their private signal than they would under anonymous scoring. Lastly, transparency can induce judges to curb their individual biases towards certain skaters; paradoxically, this may not lead to lower *aggregate* bias in the panel, as conformity concerns create the perverse incentive for judges to match the expected biases of other judges on the panel.

4. Systematic biases, especially in the form of nationalistic favoritism, has been documented in figure skating (Campbell and Galbraith, 1996; Zitzewitz, 2006; Lee, 2008; Litman and Stratmann, 2018) and other sports where performance is evaluated by judge panels (see e.g. Sandberg, 2018). Relatedly, there is evidence for home team bias and racial bias in refereeing decisions (Garicano, Palacios-Huerta, and Prendergast, 2005; Price and Wolfers, 2010; Parsons, Sulaeman, Yates, et al., 2011).

5. Thus, subjective performance evaluation has elements of a credence good (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006).

6. Indeed, committee members are frequently evaluated by comparing them to their peers. In figure skating, large deviations from average scores can lead to disciplinary actions against judges. Prendergast (1993) and Prat (2005) argue theoretically that agents may become overly conformist when the accuracy of their information cannot be perfectly verified by the principal.

Several testable predictions arise. The strongest prediction is that the dispersion of scores within the judge panel will decrease after the reform, as reputation concerns lead judges to try and match each others' scores. This conformity effect expected to be larger the more difficult it is to observe an objective score — implying in our context that conformity should be stronger for artistic elements than technical elements of the performance —, the higher public attention on the performance is, and the stronger preconceived biases, e.g. due to nationalistic favoritism, are. Perhaps surprisingly, the model predicts that nationalistic favoritism does not necessarily decrease under greater transparency, contrary to the aim of the reform. To examine the effects of the transparency reform empirically, we analyze scores from almost 17,000 figure skating performances across 127 competitions organized by the International Skating Union (ISU) between 2013 and 2020. Our empirical identification strategy compares changes in the distribution of judge scores after the 2016 transparency reform between JGP (Junior Grand Prix) events, which were not affected by the reform, and Non-JGP events, which were.

Our empirical results are broadly in line with the theoretical predictions. We find that the dispersion of scores within the judge panel for a given performance drops sharply after the reform with regard to the artistic aspects of a performance. The decrease in artistic score dispersion is both statistically significant and quantitatively sizeable, constituting approximately 9% of the pre-reform average and 29% of the pre-reform standard deviation of within-panel score dispersion, and it comes mainly from the reduction in large outliers, so judges' scores become more tightly packed around the mean. This conformity effect following the reform is absent for the technical score, which covers aspects like difficulty and execution of technical elements (jumps, spins, etc.) and is thus more objective. When investigating heterogeneity across performances, we find that the reduction in (artistic) score dispersion is particularly pronounced for high-profile events, which are likely to garner greater public attention, and for performances with a compatriot judge on the panel. Finally, we find that — contrary to the reform's original intention, but in line with our theoretical predictions — *aggregate* nationalistic bias, as measured by the average score advantage a skater receives when he or she has a compatriot judge on the panel, does not decrease significantly.

There are multiple mechanisms that can generate our results, some of which are highlighted in our theoretical framework. As we cannot determine an objective score for a performance without using the judge panel scores, we are also not able to fully distinguish between these different mechanisms empirically. However, we find neither evidence that judges give more similar scores the longer they have been evaluating together in the same panel, which speaks against social learning, nor evidence that judges anchor on previous scores or more objective aspects such as technical execution. This suggests that the decrease in score dispersion may be predominantly driven by more precise evaluations and conformity in biases, likely due to reputation concerns. Additional empirical results further support the impor-

tance of reputation concerns under the transparency regime. In line with the notion of consistency as signal of skills (Falk and Zimmermann, 2017), we document that judges' component sub-scores that add up to the overall artistic score also become more similar to each other. Furthermore, we find no evidence that the transparency reform induced a different selection of judges into committees based on observable characteristics.

Our paper contributes firstly to the literature on the consequences of transparency in committee decision-making. Theoretical models typically study how members' reputation concerns, i.e. their desire to appear competent, determine how they respond to transparency. Although transparency may under some circumstances induce anti-conformism to signal individual competence (Levy, 2007), committees may also have a preference for showing a united front in the public, in particular if true states cannot be observed ex post (Visser and Swank, 2007; Swank, Swank, and Visser, 2008; Swank and Visser, 2021). Higher transparency can also lead to more pre-decision information acquisition (Gersbach and Hahn, 2012; Swank and Visser, 2021). One difference to our setting is that these theoretical papers typically study a binary decision, whereas scores in our setting are awarded on a scale and aggregated by averaging.[7] Empirical evidence on the effect of transparency on committee decision-making is relatively scarce. Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2019) provide laboratory evidence on the role of different transparency regimes on information aggregation in groups. With regard to real-world committees, several studies examine how monetary policy deliberations responded to a reform that resulted in transcripts of FOMC meetings being made public after Fall 1993. Meade and Stasavage (2008) find that members are less likely to voice disagreement with the Committee Chairman post-reform; using computational linguistics tools, Hansen, McMahon, and Prat (2018) find that FOMC members tend to give more similar statements and engage less in back-and-forth dialogue post-reform, but also that especially rookie members seem to be better prepared with quantitative information on a diverse set of topics. Benesch, Bütler, and Hofer (2018) study a transparency reform in the Upper House of the Swiss parliament and show that, post-reform, legislators exhibit greater party discipline. Though we also find a conformity effect, there are several noteworthy differences in our setting. Firstly, the report space in our setting is continuous, which allows for strategies that do not exist under a binary report space. Secondly, and more importantly, the lack of a deliberation or discussion stage in the current setup implies that the result we find is not due to (direct) coercion or coordination with other judges. Thus, this paper thus adds to this literature by demonstrating a conformity effect under greater transparency even in the absence of information exchange, thus pro-

7. Rosar (2015) studies committee decision rules with continuous reporting and decision spaces and shows how this gives rise to incentives for strategic exaggeration.

viding stronger evidence for the way social image concerns can affect behavior of committee members.

A large number of previous studies have utilized large-scale publicly available data from professional sports contexts to investigate, among others, determinants of performance (e.g. Dohmen, 2008a; Lichter, Pestel, and Sommer, 2017), systematic decision errors (e.g. Bruine de Bruin, 2006; Pope and Schweitzer, 2011), gender differences (e.g. Böheim, Lackner, and Wagner, 2020), as well as favoritism (e.g. Garicano, Palacios-Huerta, and Prendergast, 2005; Zitzewitz, 2006; Sandberg, 2018; Fernando and George, 2021) and racial biases (e.g. Price and Wolfers, 2010; Parsons et al., 2011; Pope, Price, and Wolfers, 2018). Two closely related papers to ours are by Zitzewitz (2014) and Lee (2008), who study a set of reforms in figure skating (following a vote trading scandal at the 2002 Winter Olympics) that in fact introduced the anonymous scoring regime that was eventually reversed in 2016. Zitzewitz (2014) finds a slight but statistically insignificant increase in the compatriot score advantage after the reform, and Lee (2008) finds an increase in the standard deviation of judges' scores under anonymized publication. However, a number of other major reforms were implemented at that time, including an increase in the size of the judging panel and random dropping of judges' scores from the calculation of the final score, followed by another extensive series of reforms two years later. Our current setting using the 2016 reform allows for a cleaner attribution of changes in judge scoring behavior to increased transparency of judges' decisions, and our use of JGP events as control group in a difference-in-differences design further tightens the empirical identification by controlling for counterfactual time trends.

We also contribute to the literature studying whether changes in information structures could reduce discrimination. In recent years, a variety of reforms have been implemented at a large-scale (e.g. quotas, increased minority representation on selection committees, blind applications, pay transparency etc) to mixed results.[8] We provide a new empirical case study on the efficacy (or lack thereof) of a transparency-based method to counter favoritism/discrimination. Our results show that there is no evidence for any reduction in nationalistic favoritism following the publication of individual judge scores in figure skating. This could be due to several reasons. First, fairness norms might not be strong enough or offset by opposing loyalty norms induced by judges' home audience. Second, the group structure of committees could interact with conformity concerns, so that judges aim to give more similar scores to their peers by matching their biases, or alternatively, that the non-compatriot

8. See, e.g., Bertrand, Black, Jensen, et al. (2018) and Maida and Weber (2019) for evidence on quotas, Bagues and Esteve-Volart (2010) and Bagues, Sylos-Labini, and Zinovyeva (2017) for evidence on the effectiveness of gender representation on selection committees, Krause, Rinne, and Zimmermann (2012) and Behaghel, Crépon, and Le Barbanchon (2015) on blind applications, Mas (2017) and Baker, Halberstam, Kroft, et al. (2019) on pay transparency.

judges might skew their scores slightly upwards when one of their peers has the same nationality as the skater.[9] Third, the bias-correcting properties of aggregating multiple votes reduces the scope for reducing the aggregate bias.

The remainder of the paper is organized as follows. In Section 4.2, we describe judging in figure skating and the reform. In Section 4.3, we discuss how transparency can lead to changes in behavior through the lens of a theoretical model. We provide some descriptive statistics in Section 4.4 and outline our empirical strategy in Section 4.5. In Sections 4.6 and 4.7, we present our main results. Section 4.8 shows additional results to further explore the underlying mechanisms. Section 4.9 concludes.

4.2 Context

Background. Figure skating is a sport where athletes perform a choreographed sequence of jumps, spins, or other movements on ice to a musical track. There are four main disciplines in figure skating: Men's Singles, Women's Singles, Pairs Skating, and Ice Dance. Competitions are generally divided into two age categories: Junior events (for skaters aged 13 to 19; 21 for Dance and Pairs) and Senior events (for skaters aged 15 and above). Each season, the International Skating Union (ISU) organizes around 20 events, consisting of the European Championships, Four Continents Championships, World Championships, Olympics Winter Games, the Grand Prix Series and Final for the Senior age category; and the Junior World Championships, Junior Grand Prix (JGP) Series and Final for the Junior age category.

Performance evaluation. In a competition, skaters typically perform twice— once in the Short Program segment and once in the Free Skate segment. Within a performance, the skater executes several technical elements (i.e. jumps, spins, etc.) to a musical track, which is evaluated by a panel of 7-9 judges. The panel of judges is not supposed to confer with each other while grading the performance.[10]

During the performance, each judge assigns a Grade of Execution (ranging between -3 and +3, with increments of 1) to each technical element executed by the skater. To hinder the manipulation of scores by judges, the highest and lowest Grades are dropped. The remaining Grades are then scaled according to the difficulty of the element and averaged across judges to form the trimmed average scaled Grade,

9. Bagues and Esteve-Volart (2010) also hint at strategic dependencies between committee members leading to worse outcomes for female candidates paired with academic committees with greater female representation, as male committee members became less favorable when there were more female members on the committee.

10. The ISU states that judges should "mark independently and whilst judging [...] not converse with another Judge or indicate errors by action or sound".(International Skating Union, 2018, p. 60) However, the seating of the panel in a row means that it is possible for judges to look at the score of adjacent judges.

which is then added to a pre-determined Base Value to form the skater's points from a particular element. These are then summed up across elements to form the Technical Score.

At end of the performance, each judge assigns a score to the artistic components of performance (i.e interpretation of music, skating skills, transitions between technical elements, etc.). These artistic scores can range from 0 to 10 with increments of 0.25. As with the Technical Score, the highest and lowest component scores are dropped to compute the trimmed average of the judges' score for a particular component. Summed across components, these trimmed scores form the Artistic Score of the performance.

Note that compared to the Technical Score, judges have much more influence in determining the Artistic Score of a performance — while each technical element consists of a fixed Base Value (that cannot be influenced by the judge) and a Grade of Execution (which can be influenced by the judge), artistic components have no such fixed component, thus allowing the judge more range to favor his compatriot skater. Additionally, as opposed to the Artistic Score which is awarded at the end of the performance and for which the judge has ample time to deliberate and form beliefs regarding the scores she expects her fellow judges to award, the judge is required to assign the GOE to each Technical Element almost instantaneously, within a few seconds.

Transparency reform. After each event, detailed scoring information for all performances, including the individual scores that make up the final score, are published on the official site of the ISU. Pre-reform, with the exception of Junior Grand Prix (JGP) Series events, these individual scores were published anonymously. That is, while the identities of the judges on the panel were known, the individual scores are published in random order, so that they cannot be linked to an individual judge (see Figure 4.A.1 for an example). This lack of transparency meant that judges could not be held accountable for their decisions, which led to accusations of biased judging by the public. Such allegations came to a head with the scoring of the 2014 Olympics Women's competition, where Russian competitor Alina Zagitova was awarded gold ahead of the South Korean competitor Kim Yu-Na. Indeed, public outrage over the scoring reached such a point that the International Skating Union (ISU) considered abolishing judge anonymity in their General Meeting in 2014. While the proposal failed narrowly, it was brought up once again two years later (in 2016) and passed, so that from 2016-17 onwards, judges' scores from all competitions were published openly. Though other reforms were implemented at the 2016-17 meeting, these reforms were not explicitly aimed at reducing nationalistic judging, and mostly affect both JGP (Control) and Non-JGP (Treatment) events.[11]

11. Other reforms are mostly concerned with changes in required technical elements and updated scoring guidelines, which are typically implemented every two years (when a General Meeting is held).

Because JGP events already published scores openly prior to the transparency reform, they were unaffected by the reform and thus serve as a control group. JGP events follow the same scoring format and criteria as Non-JGP events and, to a certain extent, share the same pool of judges as Non-JGP events— over the study period of 2013-2020, half of the judges have judged in at least one JGP event and Non-JGP event. The core difference between these two groups of events lies in the level of prestige and exclusivity. JGP events are typically less prestigious and exclusive than Non-JGP events, so that scores from JGP events tend to be lower.

4.3 Potential Effects of Transparency

4.3.1 Transparency and conformity concerns

The main consequence of the transparency reform is that it increases the visibility of individual judges' evaluations to the public, allowing to better monitor them, so that social image and reputation concerns likely play a larger role in their decision-making process. The hope is that judges will want to appear competent and impartial when knowing that they might have to face the glare of public scrutiny, which in turn induces them to report performance assessment that are more accurate and less biased. Thus, judges' incentives would be better aligned with the interests of the general public.

However, in many contexts, there are no truly objective yardsticks against which to compare a judge's decision. This is blatantly clear in our figure skating setting, as the judges are hired for the very reason that there are many subjective aspects to a skating performance and that aggregating scores given by a panel of experts should approximate a fair evaluation. This is especially true for the artistic score; for the technical score, some aspects are relatively objective, e.g. not rotating fully or pre-rotating on the ice while performing a jump. Hence, a natural benchmark to compare individual judges' scores against is the overall score of the panel when aggregating all judges' scores. Outlier judges, who express very different opinions from those of their peers, may be perceived as being inattentive or biased, whereas judges who are close to the average might be perceived as competent and impartial. The transparency reform could thus create incentives to report scores that are more similar to those of others, resulting in an overall lower dispersion of scores within the panel and potentially also curbing nationalistic bias. Note that it is not possible (and not allowed) for judges to deliberate together or coordinate their actions, but judges might feel pressured to exert more effort into the accurate grading of performances, and they may collectively become more conservative, e.g. by avoiding scores that deviate too much from some common prior or implicit consensus.

A few rule changes are specific to Senior events; however, these are mostly specific to the technical elements.

4.3.2 Theoretical model of judge scoring behavior

To formalize these intuitions of how transparency can affect the distribution of scores within the judge panel, we present a theoretical model of judge scoring behavior that is based on the well-studied "beauty contest" framework introduced by Morris and Shin (2002).

4.3.2.1 Basic setup

Skater i performs in a competition. Judges $j = 1, ..., N$ sit in the panel and evaluate the quality of the performance by each reporting a score s_{ji} without joint deliberation. These individual scores $s_{1i}, ...s_{Ni}$ are then aggregated to an overall average score $s_i = \frac{1}{N}\sum_j s_{ji}$. For simplicity, we abstract from the trimming of the highest and lowest scores.

The "true" quality of the performance θ_i is drawn from a normal prior distribution with mean μ_{θ_i} and non-zero variance $\sigma^2_{\theta_i}$ and it is imperfectly observable ex post. However, by observing the performance, judges receive a noisy public signal y_i and a noisy private signal x_{ji}:

$$y_i = \theta_i + \nu_i\,, \tag{4.3.1}$$

$$x_{ji} = \theta_i + \varepsilon_{ji}\,, \tag{4.3.2}$$

where ν_i is normally distributed with mean 0 and non-zero variance $\sigma^2_{y_i}$, so the common posterior z_i is normally distributed with mean $\mu_i = E[\theta_i|y_i] = \lambda_i\,\mu_{\theta_i} + (1 - \lambda_i)\,y_i$, where $\lambda_i = \frac{\sigma^2_{y_i}}{\sigma^2_{\theta_i}+\sigma^2_{y_i}}$, and variance $\sigma^2_i = \frac{\sigma^2_{\theta_i}\sigma^2_{yi}}{\sigma^2_{\theta_i}+\sigma^2_{y_i}}$. Intuitively, the public signal can be understood as encompassing aspects of a performance that judges can easily observe and unanimously agree upon, such as obvious stumbling or falling. Each judge's private signal is unbiased but contains an idiosyncratic noise term ε_{ji} that follows a normal distribution with mean 0 and variance σ^2_i/τ_j, where $\tau_j \in (1, \infty)$ denotes the precision of judge j's signal. For example, an experienced and attentive judge may be able to evaluate the quality of a performance more reliably than a judge who is inexperienced or inattentive. As there is a strong artistic aspect to figure skating and thus no single objective criterion for evaluating a performance, it is plausible that the private signal after observing the performance is less noisy than the prior ($\tau_j > 1$), but never so precise that θ_i is perfectly observed ($\tau_j < \infty$). This offers a rationale for assigning final scores by aggregating the opinion of multiple judges in order to reduce the influence of idiosyncratic factors.

4.3.2.2 Simplified model

To lay some groundwork, we will first present a stripped-down version of our model in which judges behave non-strategically and in which signal precision τ_j is given

exogenously. We assume that judges are partially motivated to give a genuinely accurate assessment of the performance quality when reporting their scores, but they can additionally be biased towards rewarding systematically higher or lower scores to skater i. This bias may reflect favoritism, e.g. due to same nationality or a preferred skating styles (Zitzewitz, 2006; Litman and Stratmann, 2018), but it could in principle also reflect differences in judges' general strictness towards all skaters, if the bias is invariant of the skater identity. We model these two motives through the following payoff function:

$$u_j(s_{ji}, b_{ji}, \theta) \quad = \quad -(s_{ji} - \theta_i - b_{ji})^2 \,. \tag{4.3.3}$$

b_{ji} is the (fixed) bias of judge j towards skater i. Judges choose s_{ji} to maximize their expected utility. The quadratic loss formulation leads to a classical signal extraction problem, and the optimal non-strategic report $\tilde{s}_{ji}$ can be obtained using Bayes' rule:

$$\tilde{s}_{ji} \quad = \quad E[\theta_i | x_{ji}, y_i] \; + \; b_{ji} \quad = \quad \frac{1}{1+\tau_j}\,\mu_i \; + \; \frac{\tau_j}{1+\tau_j}\,x_{ji} \; + \; b_{ji} \,. \tag{4.3.4}$$

The first component $E[\theta_i|x_{ji}, y_i]$ is a linear combination of the private signal x_{ji} and the common posterior z_i and represents the actual posterior belief about performance quality θ_i that the judge forms. The more accurately a judge is able evaluate the performance, i.e. the higher τ_j, the more weight will be put on his or her actual signal. The second component b_{ji} creates a distortion in the reported score due to the judge's bias towards skater i. Depending on how the biases are distributed across judges in the panel, they may not completely average out when scores are aggregated, so some skaters may have an unfair advantage compared to others, if it so happens that the panel is tilted in favor of them, e.g., if a compatriot judge sits on the panel.

Assuming homogenous precision $\tau_j = \tau$ for all judges, the expectation and variance of scores across judges in the panel conditional on the performance θ_i are

$$E[\tilde{s}_{ji}|\theta_i] \quad = \quad \theta_i \; + \; \frac{1}{1+\tau}\,\lambda_i \,(\mu_{\theta_i} - \theta_i) \; + \; E[b_{ji}] \,, \tag{4.3.5}$$

$$Var[\tilde{s}_{ji}|\theta_i] \quad = \quad \frac{\tau}{(1+\tau)^2}\,\sigma_i^2 \; + \; Var[b_{ji}] \,. \tag{4.3.6}$$

The overall score can be ex post biased from two sources. First, the report is slanted towards the prior expectation μ_{θ_i}, because judges can only observe θ_i with noise. Hence, hypothetically, the identical performance delivered by a famous world-class skater may be awarded a higher score than if delivered by an unknown rookie skater — this is sometimes referred to as the Matthew effect (Merton, 1968; Kim and King, 2014). Second, a skater will receive systematically higher or lower scores if

there is asymmetry in judges' biases, for example if one judge exhibits strong nationalistic favoritism and the other judges in the panel are unbiased. While public focus often lies on systematic biases, a reduction in noise can be just as important to ensure the validity of a performance evaluation process (Kahneman, Sobony, and Sunstein, 2021). The expected variance in the reports given by different judges depends on the signal precision τ and the dispersion of biases within the panel. The variance decreases when performance quality can be observed more accurately, i.e. $\partial Var[\tilde{s}_{ji}|\theta_i]/\partial\tau < 0$,[12] and when there is less bias heterogeneity across judges.

4.3.2.3 Full model with conformity concerns and endogenous signal precision

Our full model extends the non-strategic setup from above with two elements. First, judges have image or reputation concerns, meaning that they want to appear competent in the way they award scores to a skating performance — both to the public and possibly also to themselves. As performance quality is not perfectly observable even ex post, especially with regard to the more artistic aspects, one straightforward way to evaluate a judges' score is to compare it to the score of other judges. Therefore, we model image concerns in a way that they lead to a motive for conforming with other judges, i.e. by not deviating too far from their scores. Second, we allow judges to endogenously adjust their signal precision τ_j by choosing their level of effort or attentiveness when observing the performance (Colombo and Femminis, 2008). Judges' payoff function is

$$u_j(s_i, \tau_j, \theta_i) \;=\; -\left(s_{ji} - \theta_i - b_{ji}\right)^2 \;-\; \eta\left(s_{ji} - \tfrac{1}{N-1}\sum_{l\neq j} s_{li}\right)^2 \;-\; C(\tau_j)\,, \qquad (4.3.7)$$

where $\eta \in (0,1)$ captures the strength of the conformity motive relative to the truthfulness motive, and $C(\tau_j)$ is the effort cost necessary to achieve precision level τ_j. Following Colombo and Femminis (2008), we assume linear costs $C(\tau_j) = c\tau_j$. The unit "price" of precision is $c \in (0, \bar{c})$, with upper limit $\bar{c} = \frac{\sigma^2}{4(1+\eta)}$ to ensure that agents choose signal precisions τ_j that are not implausibly low.[13] Note that there is now a strategic aspect to reporting behavior, since judge j's expected utility depends on the scores of the other judges, and vice versa. As a solution concept, we compute the symmetric Bayesian Nash equilibrium, in which each judge makes inferences about

12. This holds true given our assumption that $\tau > 1$. Generally, there is an inverse U-shaped relation between the score variance and τ, because when private signals are very noisy ($\tau < 1$), then judges will place a high weight on their common posterior z_i, which in turn leads to very uniform scores.

13. As we will later see, this condition on c implies that $\tau > 1 + \eta$ and ensures that the agent will always place more weight on her private signal than the common posterior when reporting the performance score. It is thus slightly stronger than the assumption $\tau > 1$ in section 4.3.2.2, although still relatively weak given that it holds for all $\tau \in [2, \infty)$.

the distribution of other judges' signals based on her own signal and then awards her optimal scores in response to other judges' reporting strategy. The individual rationality condition requires that for all $j = 1, ..., N$ and $l \neq j$,

$$\begin{aligned} s_{ji} &= \frac{1}{1+\eta}\left(E[\theta_i|x_{ji}, y_i] + b_{ji}\right) + \frac{\eta}{1+\eta} E[s_{li}|x_{ji}, y_i] \\ &= \frac{1}{1+\eta}\,\tilde{s}_{ji} + \frac{\eta}{1+\eta} E[s_{li}|x_{ji}, y_i]\,. \end{aligned} \tag{4.3.8}$$

As already observed by Morris and Shin (2002), a symmetric equilibrium implies that we can plug in the best response s_{li} from equation (4.3.8) for all $l \neq j$, leading to a feedback loop of higher-order beliefs:

$$\begin{aligned} s_{ji} &= \frac{1}{1+\eta}\,\tilde{s}_{ji} + \frac{\eta}{(1+\eta)^2}\, E[\tilde{s}_{li}|x_{ji}, y_i] + \frac{\eta^2}{(1+\eta)^3}\, E\big[E[\tilde{s}_{mi}|x_{li}, y_i]\,|\,x_{ji}\big] + \dots \\ &= \frac{1}{1+\eta}\, b_{ji} + \sum_{n=0}^{\infty} \frac{\eta^n}{(1+\eta)^{n+1}} \left[\mu_i + \frac{\tau^{n+1}}{(1+\tau)^{n+1}}\,(x_{ji} - \mu_i) + \frac{1+\eta}{\eta} E[b_i]\right] \end{aligned} \tag{4.3.9}$$

Notice that higher-order expectations about θ_i approach the common posterior z_i as n becomes large. The geometric series in equation (4.3.9) is bounded and converges to a unique social equilibrium in which every judge j reports

$$s_{ji} = \frac{1+\eta}{1+\eta+\tau}\,\mu_i + \frac{\tau}{1+\eta+\tau}\,x_{ji} + \frac{1}{1+\eta}\,b_{ji} + \frac{\eta}{1+\eta} E[b_i]\,. \tag{4.3.10}$$

This equilibrium condition has to be true regardless of the level of precision τ that judges choose. Holding constant τ, the optimal strategic report s_{ji} is attenuated more strongly towards the common posterior z_i compared to the non-strategic report $\tilde{s}_{ji}$. Hence, it resembles a tacit coordination of judges to deviate from their true posterior beliefs of performance quality and move their scores closer towards an uncontroversial benchmark. Interestingly, the desire to appear more in line with other judges also leads to higher conformity in biases, as judges now realign their bias partially towards the expected bias $E[b_i]$.

We now move on to find the equilibrium level of effort τ. Let all judges $l \neq j$ follow the same strategy with report s_{ji} from equation (4.3.10) and effort level $\tau_l = \tau$. Judge j takes this as given and seeks to find the optimal individual effort level τ_j. Her best response in equation (4.3.8) then becomes

$$s_{ji}(\tau_j) = \frac{1+\tau+\eta(1+\tau_j)}{(1+\tau_j)(1+\eta+\tau)}\,\mu + \frac{\tau_j\,(1+\tau)}{(1+\tau_j)(1+\eta+\tau)}\,x_{ji}\,. \tag{4.3.11}$$

We can plug this into the expected utility of j and derive the first-order condition (FOC) with regard to τ_j, yielding

$$\frac{\partial}{\partial \tau_j} E[u_j(s_i, \tau_j, \theta_i)] \;=\; \frac{(1+\eta)(1+\tau)^2}{(1+\tau_j)^2(1+\eta+\tau)^2}\,\sigma^2 \;-\; c \;\stackrel{!}{=}\; 0 \tag{4.3.12}$$

As the problem is concave in τ_j, the necessary FOC is also sufficient. In a symmetric equilibrium, all judges j choose the same level of effort that fulfills the FOC in equation (4.3.12), hence it must hold that the optimal signal precision is

$$\tau_j \;=\; \tau \;=\; \sqrt{1+\eta}\cdot\frac{\sigma}{\sqrt{c}} \;-\; (1+\eta)\,. \tag{4.3.13}$$

This term is increasing in the conformity concern η for all $c \in (0, \bar{c}]$. Hence, transparency can be used as incentive mechanism for inducing higher judge effort when evaluating skater performances.

Conditional on θ_i, the expectation and variance of performance scores across judges look as follows when taking into account conformity concerns and endogenous signal precision:

$$\begin{aligned} E[s_{ji}|\theta_i] \;&=\; \theta_i \;+\; \frac{1+\eta}{1+\eta+\tau}\,\lambda_i\,(\mu_{\theta_i} - \theta_i) \;+\; E[b_{ji}] \\ &=\; \theta_i \;+\; \frac{\sqrt{(1+\eta)\,c}}{\sigma_i}\,\lambda_i\,(\mu_{\theta_i} - \theta_i) \;+\; E[b_{ji}]\,, \end{aligned} \tag{4.3.14}$$

and

$$\begin{aligned} Var[s_{ji}|\theta_i] \;&=\; \frac{\tau}{(1+\eta+\tau)^2}\,\sigma_i^2 \;+\; \frac{1}{(1+\eta)^2}\,Var[b_{ji}] \\ &=\; \frac{\sqrt{c}\,\sigma}{2\,(1+\eta)^{\frac{3}{2}}} \;-\; c \;+\; \frac{1}{(1+\eta)^2}\,Var[b_{ji}] \end{aligned} \tag{4.3.15}$$

4.3.2.4 Predicted effects of the transparency reform

Under anonymous scoring, the public cannot observe which judge gave which score. Hence, judges do not have to worry much about appearing incompetent or biased when the score they award is discrepant from the other judges' scores. In contrast, when scoring becomes transparent, judges may start worrying more about their social image and their desire to appear competent. We therefore interpret the transparency reform as an increase in η in our model. This allows us to derive a number of predictions for how the distribution of scores across judges changes.

(1) Lower score dispersion for a given performance. — Inducing stronger conformity concerns leads to a lower variance of judge scores for any given performance:

$$\frac{\partial}{\partial \eta} Var[s_{ji}|\theta_i] \;<\; 0\,. \tag{4.3.16}$$

This decrease in score dispersion results from three sources. First, stronger conformity concerns result in scores that are more conservative in the sense that they are attenuated towards the prior z_i, which means that judges place less weight on their idiosyncratic signals. Second, increasing effort in η leads to less noise in judges' private signals. Third, dispersion can further decrease due to judges adjusting their individual biases more towards the average bias in the panel, which implies that the impact of transparency may be stronger for performances in which judges are very polarized in their biases toward the skater.

(2) Effect on score dispersion increases in subjectivity. — Skaters are evaluated both on the technical aspects and the artistic aspects of their performance. The latter is arguably much more subjective than the former, which implies that judges may have a harder time trying to award the artistic score as accurately as possible. We therefore look at another comparative static, which is how the effect of transparency on dispersion of scores is affected by an increase in the cost c of obtaining a more precise signal of performance quality. It is straightforward to show that

$$\frac{\partial^2}{\partial \eta\, \partial c} Var[s_{ji}|\theta_i] \;<\; 0\,. \tag{4.3.17}$$

This implies that the reduction in score dispersion in prediction (1) is more pronounced if objective performance evaluation is more difficult. In particular, we would expect to see a larger reduction in dispersion for the artistic score than for the technical score.

Further rationales for expecting smaller effects for the technical score is that conformity to other judges may play less of a role (i.e. η is lower), because its relative objectivity makes it more important for reputation-concerned judges to give their most accurate assessment, or because technical scores are awarded almost instantaneously and judges may not have time to consider other judges' behavior.

(3) No decrease in aggregate bias. — Interestingly, our model suggests that, on average, higher transparency may leave the *aggregate* bias $B_i = \sum_j b_{ji}$ of the panel towards skater i unchanged, as the bias component in equation (4.3.14) is invariant to η:

$$\frac{\partial^2}{\partial \eta\, \partial E[b_{ji}]} E[s_{ji}|\theta_i] \;=\; 0\,. \tag{4.3.18}$$

The reason is that with conformity concerns, judges also incorporate beliefs about other judges' biases $E[b_i]$ in order to match their scores more closely. This prediction is consistent with the results in Sandberg (2018), who finds that judges in dressage competitions favor athletes of the same nationality as other judges on the panel. In our context, one may therefore also expect conformity effects to be particularly strong when judge biases can be easily inferred, such as when there are matching nationalities.

4.3.3 Further potential effects of transparency on judge behavior

Transparency may also affect judge behavior through other mechanisms that are not explicitly included in our model. In the following, we will briefly discuss some of these mechanisms and how they may affect our theoretical predictions.

Consistency as a signal of skill. Apart from trying to report scores that are closer to the those of other judges on the panel, judges could also try to signal their competence to the public in the absence of objectively verifiable yardsticks by being more consistent in their evaluations (Falk and Zimmermann, 2017). For example, the overall artistic score is composed of scores in several sub-components, and large discrepancies in a judge's scores across these different artistic components of a performance might be perceived as arbitrary scoring behavior, and more uniform evaluations may be interpreted as confidence in one's assessment, regardless of whether that is true or not. As a consequence, judges may want to report more consistent scores for each score component after the transparency reform.

Appealing to the home constituency. Judges may want to appear competent and impartial when their evaluations are potentially subject to public monitoring. But more generally, social image and reputational concerns tend to induce individuals to act in a way that is in accordance with the prevailing norms and expectations that they face, which might partially contradict each other. For example, audiences in the judge's home countries as well as the national federation that appointed the judge may in fact expect him or her to favor compatriot skaters and discriminate against rival skaters (Zitzewitz, 2006). Hence, it is also possible that making individual scores transparent instead exacerbates nationalistic judging.[14] In this case, judges' scores would become more polarized, as judges try to signal loyalty to their

14. Dohmen (2008b), for instance, finds that football referees exhibit home team favoritism, in particular when the physical distance of the public crowd to the field is smaller, and when the crowd consists of supporters of the home team. Benesch, Bütler, and Hofer (2018) find greater party discipline after the transparency reform in the Swiss Upper House, even though this is not necessarily in line with the preferences of the median cantonal voter. Stasavage (2007) finds that in a model with biased and unbiased experts, unbiased experts only vote truthfully under public voting if reputational concerns are sufficiently weak.

respective constituencies. Importantly, this would predict an *increase* in nationalistic bias and in score dispersion following the transparency reform, in particular for performances with a compatriot judge on the panel.

Exaggeration and counter-exaggeration. When there is a potentially biased judge on the panel, other judges can in fact react to this strategically by biasing their scores in the opposite direction if they have fairness concerns for the aggregate score awarded to skaters (Li, Rosen, and Suen, 2001; Rausser, Simon, and Zhao, 2015). Transparency could potentially break such feedback loops of bias and counterbias, which would also predict a decrease in score dispersion for a given performance, though mostly concentrated on performances where the presumed biases are particularly strong, e.g. when there is a compatriot judge on the panel. Note, however, that some previous studies on the behavior of sports judge panels find that non-compatriot judges may in fact move their scores closer towards those of the compatriot judge instead of in the opposite direction (Zitzewitz, 2006; Sandberg, 2018).

Vote trading and rigging. Transparency can also facilitate corruption, e.g. by rigging or vote trading, because potential bribers can now verify whether the bribed judge actually followed through, and colluding judges can better monitor each others' behavior and implement repeated game strategies.[15] In fact, anonymous voting was first introduced by the ISU in 2002 precisely in response to a vote trading scandal at the Salt Lake City Olympics. However, it is difficult to predict in which ways increased collusion or results fixing would affect the observed scoring patterns, assuming that due to public scrutiny, vote trading strategies need to sophisticated enough that they are not easily detectable.[16] Since outright collusion and cheating are risky endeavors and the success chances are uncertain given that each individual judges only has a limited impact on total scores, we would likely not observe strong universal changes in judging behavior due to fixing alone.

4.4 Data and Descriptive Statistics

To see how the reform affected panel scoring, we obtain from the ISU website information on skaters' performances at all official ISU competitions in the 2013-14 sea-

15. For example, Edward J. Green and Robert H. Porter (1984) show in the context of firm cartels that it can be easier to uphold collusion when public monitoring is possible.

16. For example, in the 2002 scandal, a French judge admitted (though later recanted) to having been pressured by her national federation to rank the Russian pair first in the pairs' competition, in exchange for higher votes to a French couple that would perform in the ice dance competition a couple of days later.

son to the 2019-20 season, three seasons pre-reform and four seasons post-reform.[17] This information includes all scores given by judges on the panel towards each technical element and artistic component of the performance, as well as the identities and nationalities of the skater and the judges on the panel. We restrict the dataset to performances from competitions where there is a full panel (9 judges).[18] Table 4.4.1 presents descriptive statistics for the dataset. In total, the dataset comprises 16,821 observations (performances) across 127 events and 1,028 rounds (competition $\times$ segments).[19] These performances are given by 1,905 skaters, and judged by 611 judges.

There are three main differences between the JGP (Control) and Non-JGP (Treatment) groups. Firstly, there are fewer JGP (Control) events (7 each season), so that there are fewer JGP (Control) performances in general. Secondly, compared to Non-JGP (Treat) performances, Artistic Scores are lower in the JGP (Control) group (JGP pre-reform: 24.77; Non-JGP pre-reform: 33.87), reflecting the lower prestige of JGP (Control) events, and hence the overall lower quality of performances. Thirdly, the proportion of performances where the skater has a compatriot judge on the panel is higher for Non-JGP (Treat) events, likely because countries such as China, Russia, US and Japan, where figure skating has traditionally been a national sport, are more likely to have judges sufficiently qualified to judge in Non-JGP events, as well as the higher number of skilled skaters from these nations.

In general, Compatriot performances tend to be scored more highly compared to Non-Compatriot performances, though this could be due to the higher skill level of skaters from countries with higher likelihood of serving on panels. Furthermore, this difference is higher for Non-JGP (Treat) events, perhaps reflecting the higher stakes nature of these events, such that judges have stronger incentives to favor their compatriot skaters.

Our main measure of score dispersion is the within-performance standard deviation (Panel StD), which is the standard deviation of scores given by the panel of judges to a performance, and is computed as $\sigma_{sp} = \sqrt{(\frac{1}{9})\sum_{j=1}^{9}(\pi_{spj} - \bar{\pi}_{sp})^2}$, where

17. Though data is available until the 2005-06 season, the main presented results are restricted to observations from the 2013-14 season onwards. This is firstly due to a number of changes in event formats in the 2010-11 and 2011-12 seasons (e.g. the Compulsory Dance and Original Dance segments were replaced with the Short Dance segment; instead of holding a Preliminary Qualification Round in Senior events, qualifications were done based on scores from the Short Program after the 2011-12 season.), so that it is not possible to control for discipline $\times$ segment. Secondly, JGP (Control) skaters typically do not have long careers, so these skaters are no longer in the dataset after a few years; results with skater FEs are mainly identified from performances close to the reform period. Results using the full dataset (without skater FEs or discipline $\times$ segment controls) are presented in the Appendix.

18. Due to budget constraints, some competitions (typically JGP (Control)) have panels with fewer than 9 judges (5-8 judges). Nonetheless, such panels are uncommon, consisting only of 520 observations. Including these performances does not lead to in any significant changes in results.

19. For example, the 2014 World Figure Skating Championships constitutes an event. Within this event, there are 4 competitions, one for each discipline, and 8 rounds. Variation in the panel of judges occurs at the round level.

Table 4.4.1. Descriptive Statistics

	All Obs.	Non-JGP Pre	Non-JGP Post	JGP Pre	JGP Post
# Performances	16,821	3,994	5,384	3,103	4,340
# Skaters	1,905	617	730	711	954
# Judges	611	328	351	342	392
# Events	127	34	44	21	28
# Rounds	1028	292	384	152	200
Proportion Compatriot Performances	0.61	0.66	0.68	0.54	0.52
Mean Trimmed Total Score	76.75	82.73	89.67	61.42	66.19
Mean Trimmed Technical Score	39.16	42.08	46.04	31.09	33.72
Mean Trimmed Artistic Score	31.06	33.87	36.23	24.77	26.55
Mean Panel StD in Total Score	3.13	3.20	3.17	2.98	3.13
Mean Panel StD in Technical Score	1.33	1.40	1.56	1.03	1.18
Mean Panel StD in Artistic Score	1.75	1.78	1.62	1.83	1.84
Total Score (Compatriot)	81.62	86.42	94.04	63.18	69.52
Technical Score (Compatriot)	41.61	43.78	48.16	32.02	35.56
Artistic Score (Compatriot)	33.08	35.29	37.83	25.92	28.08
Total Score (Non-Compatriot)	69.26	75.52	80.54	59.39	62.61
Technical Score (Non-Compatriot)	35.39	38.75	41.61	30.02	31.74
Artistic Score (Non-Compatriot)	27.95	31.09	32.89	23.43	24.91

Notes: Descriptive statistics for dataset with all observations.

π_{spj} is the score given by judge j towards performance p by skater s. From Table 4.4.1, it can be seen that, pre-reform, the panel standard deviation does not differ much between Non-JGP (Treat) and JGP (Control) performances– it is 1.78 for Non-JGP (Treat) performances, and 1.83 for JGP (Control) performances (KS-test p-value = 1). This is likely due to the same scoring system used in both types of events, as well as the partial overlap in judges and skaters in both types of events. Overall, scoring seems to reflect performance quality, but within-performance scoring behavior is similar between JGP (Control) and Non-JGP (Treat) performances.

4.5 Empirical Strategy

4.5.1 Identification

We use a difference-in-differences approach to empirically identify the effects of the transparency reform on judges' performance evaluation behavior, using performances in JGP events as control group, since deanonymized scores were already published before the 2016 reform for these events. The main identification assumption is that performance scores in Non-JGP events and in JGP events would have followed the same counterfactual time trend in absence of the transparency reform. While JGP events are notably less prestigious than Non-JGP events, so the average

quality of performances is also lower, any level differences in score statistics between these events are not problematic as long as the common trends assumption holds. Moreover, we assume that the reform does not affect skaters' performance per se, but only the way judges award scores for these performances. This seems plausible given that for skaters, nothing changes about how and when they learn about the scores they receive for their performance.

Ideally, we would study deanonymized judge scores both before and after the reform, for example to evaluate how behavior changes for a compatriot judge on the panel compared to non-compatriot judges, or how the same judge behaves under different publication policies. Unfortunately, it is precisely the anonymization of individual judges' scores that prevents any analyses that require scores to be matched to judge identity before the reform. Therefore, we will mainly look at judge panel-level statistics such as the aggregate score or the within-panel score dispersion as outcome variables. This means that we are not able to identify the extent of favoritism by the compatriot judge him-/herself prior to the reform for Non-JGP events. Instead, we will investigate the *aggregate* net bias of a skaters' score when there is a compatriot judge on the panel, which may also include potential favoritism, e.g. due to bloc-voting, or strategic (counter-)exaggerations by non-compatriot judges.

4.5.2 Estimation

In our baseline specification, we estimate the following difference-in-differences model using judge score data at the performance-level:

$$y_{isp} = \alpha_i + \beta_1 \cdot NonJGP_{isp} + \beta_2 \cdot NonJGP \times Post_{isp} + \delta_s + \varepsilon_{isp}\,, \qquad (4.5.1)$$

where y_{isp} is the outcome variable for performance p by skater i in season s. $NonJGP_{isp}$ is an indicator variable for performances at Non-JGP events. δ_s represents season fixed effects that capture any changes in score statistics over time. The main independent variable of interest is $NonJGP \times Post_{isp}$, which is the interaction of the Non-JGP indicator with an indicator for post-reform events (season 2016-17 onwards). Hence, β_2 is the estimated treatment effect of the transparency reform on y_{isp}. For further robustness, we also estimate additional specifications adding skater fixed effects α_i. We will mainly use the regression model in equation 4.5.1 to estimate effects on the score dispersion within a panel for a given performance, as measured by the standard deviation $\sigma_{isp} = \sqrt{\frac{1}{9}\sum_{j=1}^{9}(\pi_{jp} - \bar{\pi}_{jp})^2}$, where π_{jp} is the score awarded by judge $j = 1, ..., 9$ and $\bar{\pi}_{jp}$ is the (untrimmed) average score by all judges.

To estimate the net degree of nationalistic favoritism and how it is affected by the transparency reform, we compare the aggregate scores for performances by skaters with a compatriot judge on the panel with scores for similar performances by skaters whose nation is not represented on the judge panel. This additional variation within the same round now allows us to estimate the following model:

$$\begin{aligned} y_{irp} \;=\; & \alpha_i + \beta_1 \cdot \mathit{Comp}_{irp} + \beta_2 \cdot \mathit{Comp} \times \mathit{NonJGP}_{irp} \\ & + \beta_3 \cdot \mathit{Comp} \times \mathit{Post}_{irp} + \beta_4 \cdot \mathit{Comp} \times \mathit{NonJGP} \times \mathit{Post}_{irp} \\ & + \delta_r \;+\; \varepsilon_{irp}\,, \end{aligned} \tag{4.5.2}$$

where Comp_{irp} indicates if skater i has a compatriot judge on the panel for his or her performance p in round r. Therefore, if y_{irp} is the score a skater receives for his or her performance, then β_1 represents the baseline degree of nationalistic bias in figure skating scores. We further interact the compatriot performance indicator with an indicator for Non-JGP events ($\mathit{Comp} \times \mathit{NonJGP}_{irp}$), to control for time-invariant differences between the level of favoritism between JGP and Non-JGP events, and with an indicator for post-reform events ($\mathit{Comp} \times \mathit{Post}_{irp}$), to control for common changes over time. Finally, the triple-interaction term $\mathit{Comp} \times \mathit{NonJGP} \times \mathit{Post}_{irp}$ now allows us to estimate how the transparency reform affects the compatriot score advantage based on a difference-in-differences approach. To ensure the comparability of performances, we include round fixed effects δ_r, so that nationalistic bias is always estimated using score differentials of skaters with and without compatriot judges that perform in the same round. Nevertheless, while the exact composition of the panel can be treated as random from a skater's point of view, it is not necessarily random whether there is a compatriot judge on it or not. This is because countries with traditionally strong figure skating athletes also tend to be overrepresented in judge panels, since judges themselves are usually former professional skaters. Therefore, we further include skater fixed effects α_i in our regression to control for differences in skater ability.[20]

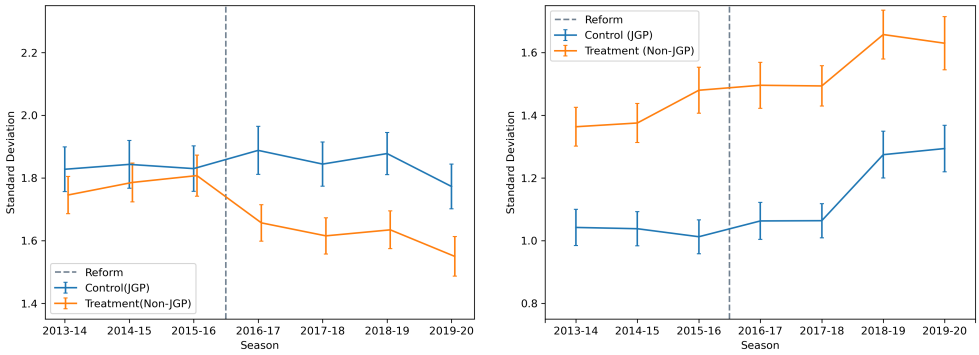
4.6 Effects on Score Dispersion

4.6.1 Average score dispersion across judges

First, we examine whether the transparency reform affected the dispersion of scores across judges for the same performance. Figure 4.6.1 plots the average season-by-season within-panel standard deviations of the artistic score and the technical score, respectively, separately for Non-JGP and JGP performances. Reassuringly, the panel standard deviations for Non-JGP and JGP performances essentially follow a nice parallel trend prior to the reform.[21] Strikingly, there is a sharp drop in the artistic score

20. Note, however, that controlling for measures of skater quality may affect the implicit weights of observations when identifying the compatriot score advantage, as for some skaters we observe few or no performances without a compatriot judges on the panel, and the number and composition of these skaters may vary over time and events.

21. To further examine the plausibility of the parallel trend assumption, we plot in the Figure 4.A.2 season-by-season panel standard deviations (as in Figure 4.6.1), but with an extended pre-reform period, starting from the 2005-06 season, which is the first season under the current ISU scoring system.



(a) SD of artistic score **(b)** SD of technical score

Figure 4.6.1. Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2013-14 to 2017-18.

Notes: Each blue (orange) point indicates the average panel standard deviation for a season, for JGP (Control) and Treatment (Non-JGP) performances. The dashed line indicates the implementation of the transparency reform; error bars are 95% confidence intervals.

dispersion for Non-JGP performances after 2016, whereas the dispersion remains about constant for JGP (Control) performances, and this gap persists for all the remaining seasons that we observe. This provides some first descriptive evidence that judges within a panel award more similar scores for the same performance under transparency than under anonymity. However, there is no such effect for the technical score. While the pre-reform gap between treatment and control performances is much starker, the difference remains more or less constant post-reform. The general increase in the technical score dispersion from season 2018-19 onwards is likely due to a scoring reform that increases the range of possible GOEs that judges can assign from 7 points (-3 to 3 in one-point increments) to 11 points (-5 to 5 in one-point increments).

Table 4.6.1 presents the formal difference-in-differences estimates based on regression equation 4.5.1 and generally confirms the pattern we observe in Figure 4.6.1. The main coefficient of interest here is $Post \times Non - JGP$, which is an indicator for treated events after the transparency reform. Columns (1) show that this coefficient is negative and highly significant for the artistic score, showing that judges behave more similar to each other in response to the reform. It is also robust to the inclusion of skater fixed effects in column (2). The effect size of -0.163 ($p = 0.0010$) is quantitatively meaningful, corresponding to a 9% decrease in panel standard deviation relative to the pre-reform mean of 1.778 for Non-JGP performances, or 28% of a standard deviation (across performances) in within-performance score dispersion.

In contrast, columns (3) to (5) show that there is no effect on the within-performance standard deviation of the technical elements score. While the estimates confirm that baseline score dispersion is higher for Non-JGP events, there is no decrease relative to JGP events following the reform. The coefficient of -0.048 in the specification with skater fixed effects is statistically insignificant ($p = 0.128$) and

Table 4.6.1. Effect of de-anonymized publication on standard deviation of panel scores.

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.003	-0.023	0.031	-0.011	-0.008
	(0.041)	(0.045)	(0.020)	(0.021)	(0.020)
Post × Non-JGP	-0.143***	-0.121**	-0.024	-0.037	-0.010
	(0.045)	(0.049)	(0.029)	(0.028)	(0.029)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.128	0.298	0.550	0.615	0.616

Coefficients obtained from estimating equation 4.5.1, with standard deviation of panel scores as dependent variable. Standard errors are clustered at event level. Column (5) excludes the 18-19 and 19-20 seasons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

also quantitatively smaller, corresponding to a 3.4% decrease in panel standard deviation relative to the pre-reform mean of 1.405 for Non-JGP events. Due to the change in grading scales for the technical score starting from season 2018-19, we further provide estimates that only include performances until season 2017-18 (see column (5)), and if anything, the coefficient moves closer to zero.

Overall, these estimates show that judges in a panel give more similar evaluations with regard to the artistic aspects of a performance in response to the transparency reform, but not with regard to the more technical aspects. This is in line with what our theoretical framework predicts, as social image and reputational concerns can generate an incentive to award scores that are more similar to those of other judges in the absence of objective standards against which the public can gauge the accuracy of a judge's scores; and this is particularly true for the inherently more subjective artistic score. For the more objective technical score, judges' incentives under transparency might be more tilted towards reporting their true assessment instead of what they believe others will report.

In principle, the decrease in standard deviation of artistic scores could also stem from judges behaving in a more polarized manner, in the sense that judges sort into distinct groups with high intra-group conformity but large between-group differences.[22] This could be the case if, for example, the compatriot judge (and his conspirators) feels pressured into awarding higher scores post-reform, with other

22. The standard deviation of scores could decrease with more polarized scoring if, for example, the degree of uniformity amongst the majority of scores is high. An example of this would be: In performance A, scores are [4,4,4,4,4,4,4,10,10] and in performance B, [4,4,5,5,9,9,10,10,10]. Scores are more polarized in performance A, but the overall standard deviation is higher in performance B.

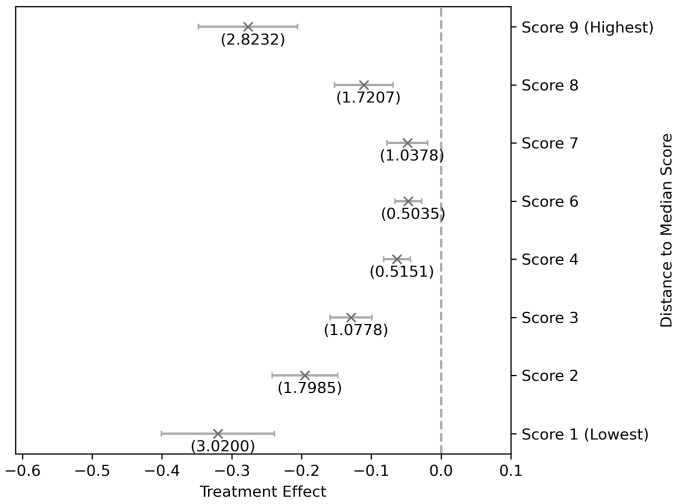


Figure 4.6.2. Estimates of Non-JGP×Post, with distance of sorted scores to median score as dependent variables

Notes: Each point plots the coefficient on Non-JGP×Post, obtained from estimating equation 4.5.1 with the distance of the k-th highest(lowest) score on the panel to the median score as the dependent variable. Controls for discipline×segment, starting order, and home event are included. Error bars indicate 95% confidence intervals; figures in parentheses indicate pre-reform means for Non-JGP (Treat) performances.

judges reacting to this by awarding lower scores, so that overall, scores become more bunched at the maximum and minimum of the panel scores. However, when we look at the distance between individual judges' scores (ordered from lowest to highest) and the median score in the panel as outcome variable, we can see that the decrease in dispersion post-reform indeed comes from all judges moving closer to the median, in particular those with the extreme scores.[23] More specifically, we use these score differences as dependent variables to estimate 4.5.1 and plot the respective coefficient estimates on Post×Non-JGP in Figure 4.6.2.

This scoring pattern suggests that, in light of the greater public visibility post-reform, judges award more similar scores, either due to an aversion of appearing out-of-line with their fellow judges or greater effort exertion to appear competent. The overall movement to the median suggests that the increased uniformity in scores unlikely to be entirely caused by judges feeling increased pressure to signal loyalty to their home countries by awarding even higher scores to their compatriot skaters, and other judges counteracting this bias by awarding lower scores.

23. Although the mean could also be used as a measure of central tendency, we opt instead for the median, so that the treatment effect captures only the combined effect of the reform on each score and the median score. Using the panel mean would capture the combined effect of the reform on each score and all other scores.

4.6.2 Heterogeneous effects

If the effects of transparency on conformity within judge panels partly work through social image and reputation concerns, then this should be particularly pronounced for competitions that generate large public attention, e.g. highly prestiguous events such as Olympics or Grand Prix competitions, at which world-class athletes perform. To analyze this, we use the average world rank of skater's performing in round *r* as proxy for the level of public attention and check for heterogeneity of effects along this dimension.[24] For this, we estimate the following regression equation:

$$\begin{aligned} \sigma_{isp} \;=\; & \alpha + \beta_1 \cdot \mathit{NonJGP}_{isp} + \beta_2 \cdot \mathit{NonJGP} \times \mathit{Post}_{isp} \\ & + \gamma_1 \cdot \mathit{RoundQ} \times \mathit{NonJGP}_{isp} \\ & + \gamma_2 \cdot \mathit{RoundQ} \times \mathit{NonJGP} \times \mathit{Post}_{isp} + \delta_s + \varepsilon_{isp}\,, \end{aligned} \tag{4.6.1}$$

where *RoundQ* is our measure for round quality, computed using the average rank of skaters performing in the round and, for ease of interpretation, standardized to mean 0 and standard deviation 1 for Non-JGP events. We extend the baseline model introduced in section 4.5 by adding interactions of the Non-JGP indicator and the post-reform Non-JGP indicator, respectively, with our round quality measure. The main coefficient of interest here is γ_2, which measures how much the treatment effect of transparency on within-panel score dispersion changes for a one standard deviation increase in round quality, which in turn can be regarded as proxy for public interest. We notably omit the main effects for *RoundQ*, so this is not a full triple-differences model. The reason for this is that JGP events, which serve as our control group, are less exclusive and prestigious than Non-JGP events as a general rule; hence it is likely that the effect of higher round quality is not comparable for these classes of events.

Table 4.6.2 presents our results on treatment effect heterogeneity for both the within-panel dispersion of artistic scores and of technical scores. We can see from columns (1) and (2) that higher average skater rank in the round indeed leads to stronger conformity in judges' artistic scores in response to the transparency reform. Our favored specification including skater fixed effects shows that a one standard deviation in round quality is associated with an additional reduction of the within-panel standard deviation by 0.076 points post-reform, which is both highly statistically significant ($p = 0.0022$) and quantitatively meaning. We find no such pattern with regard to the technical score.

24. These rankings are updated by the ISU after every event, and are computed based on the skater's highest/second highest placements at various sanctioned competitions from the previous two seasons and the current season. Some skaters are not ranked, because they placed too low in previous competitions or because they are new. See Communication No. 1629 (International Skating Union, 2010) for details regarding rank point distributions. We treat all unranked skaters as though they were one rank below the lowest-ranking skater in the respective discipline and season.

Table 4.6.2. Heterogeneous effects by average quality of skaters in a round.

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.003	0.053	0.050**	0.004	-0.003
	(0.041)	(0.044)	(0.020)	(0.028)	(0.025)
Post × Non-JGP	-0.213***	-0.251***	-0.007	-0.030	-0.017
	(0.046)	(0.048)	(0.029)	(0.033)	(0.034)
Round quality × Non-JGP	0.028**	0.061***	0.010	0.001	-0.003
	(0.013)	(0.013)	(0.009)	(0.010)	(0.011)
Round quality × Non-JGP × Post	-0.069***	-0.089***	0.012	0.005	-0.007
	(0.019)	(0.019)	(0.012)	(0.013)	(0.014)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.109	0.296	0.558	0.619	0.617

Coefficients obtained from estimating 4.5.1, with standard deviation of panel scores as dependent variable. Standard errors in parentheses are clustered at event level. Column (5) excludes the 18-19 and 19-20 seasons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Next, we investigate if there are heterogeneous effects of the transparency reform within rounds, focusing on two main aspects— the presence of a compatriot judge on the panel, and the skater's relative rank within the round. Firstly, the reduction in score dispersion following the reform could differ by compatriot and non-compatriot performances if judges' scores incorporate strategic exaggeration and counter-exaggeration motives, such as fair-minded judges wanting to compensate for favoritism by a compatriot judge on the panel. Transparency could mitigate such motives, in which case we would observe an even larger drop in the standard deviation of scores of performances with a compatriot judge on the panel. In addition, both compatriot performances and high-profile performances may attract more intense monitoring from the public. To proxy for within-round performance profile, we compute the relative world rank of the skater compared to other skaters in the round and normalize it by the number of starters — our variable for "relative rank" thus ranges from 0 (lowest-ranking skater in round) to 1 (highest-ranking skater in round).

Table 4.6.3 present the results from fixed effects regressions of within-panel standard deviation on interactions between the treatment status dummies and an indicator for compatriot performances as well as with skaters' relative world rank in the round. We always include round fixed effects, so all comparisons are between different skaters performing in the same round. For the artistic score, we find some weak evidence to support our hypotheses that scores for compatriot performances and higher-ranked skaters become more uniform in response to the transparency

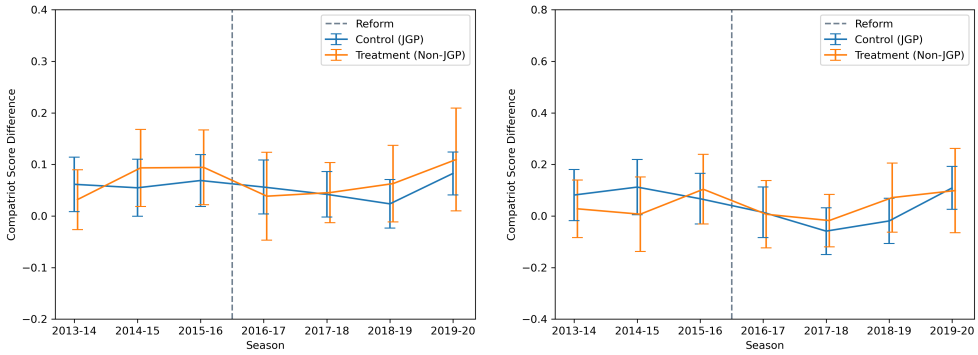
Table 4.6.3. Heterogeneous effects within rounds

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Compatriot	0.039	0.018	0.058***	0.033*	0.031*
	(0.026)	(0.031)	(0.018)	(0.018)	(0.018)
Compatriot × Non-JGP	0.035	0.062	0.011	0.008	0.006
	(0.036)	(0.038)	(0.031)	(0.026)	(0.026)
Compatriot × Post	-0.005	0.025	0.042	0.005	-0.001
	(0.033)	(0.040)	(0.028)	(0.022)	(0.021)
Compatriot × Post × Non-JGP	-0.048	-0.089*	-0.027	-0.000	-0.004
	(0.047)	(0.050)	(0.043)	(0.034)	(0.035)
Relative rank	0.110**	-0.041	0.626***	0.064*	0.012
	(0.055)	(0.055)	(0.025)	(0.037)	(0.038)
Relative rank × Non-JGP	-0.339***	-0.000	-0.190***	0.063	0.069
	(0.077)	(0.065)	(0.049)	(0.047)	(0.048)
Relative rank × Post	-0.084	-0.048	0.046	0.048	0.047
	(0.067)	(0.069)	(0.045)	(0.049)	(0.058)
Relative rank × Non-JGP × Post	-0.037	-0.131	-0.122	-0.126*	-0.094
	(0.098)	(0.090)	(0.075)	(0.065)	(0.082)
Skater FEs	—	Yes	—	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.290	0.439	0.542	0.650	0.646

Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

reform. The point estimates for both the compatriot and the rank triple-interaction are always negative, indicating an additional drop in score dispersion, although statistically insignificant in column (1). After including skater fixed effects in column (2), the compatriot interaction becomes significant at the 10% level and is quantitatively about half of the original treatment effect coefficient in table 4.6.1. The interaction with relative rank remains insignificant, although the point estimate suggests that the decrease in panel standard deviation is 0.12 points larger for the highest-ranking skater compared to the lowest-ranking skater per round. For the technical score, we find no evidence for heterogeneity by compatriot versus non-compatriot performances. Although there is no decrease in within-panel score dispersion on average, as we reported in section 4.6, there is some weak indication that the technical scores for higher-ranked skaters in a round become more similar, but column (5) shows that this might be driven mostly by the change in GOE scale that became active in 2018.

Overall, we find patterns of heterogeneity in this section that is consistent with the hypothesis that the higher degree of conformity, in the form of lower dispersion



(a) Artistic score **(b)** Technical score

Figure 4.7.1. Compatriot score advantage for JGP (Control) and Non-JGP (Treat) events, from seasons 2013-14 to 2019-20.

Notes: Each blue (orange) point indicates the average within-round compatriot score differential by season, separately for JGP (Control) (Treatment(Non-JGP)) performances. The dashed line indicates the implementation of the transparency reform; error bars are 95% confidence intervals.

of (artistic) scores within the panel, is driven by stronger social image concerns when each judge's score is published openly. The social image channel predicts that the transparency reform should have stronger effects for performances that receive higher public interest, and this is precisely what we find.

4.7 Effects on Nationalistic Bias

4.7.1 Effect on average compatriot score advantage

We next look at how the transparency reform affected nationalistic favoritism, as measured by how large the score advantage is for skaters with a compatriot judge on the panel compared to similar skaters without a compatriot judge on the panel. To make the outcome more comparable across rounds, we rescale each skater's scores such that one unit corresponds to the standard deviation of scores across skaters within the respective round. Intuitively, the problem is that even a small positive bias in a skater's scores can result in a sizeable advantage for the final ranking when the scores of all competitors are very close to each other, whereas it would be of little consequence when the competitors' scores are far apart from each other anyway.

As first descriptive evidence, Figure 4.7.1 plots the evolution of (within-round) compatriot score differentials over time, separately for JGP and Non-JGP events. First, we observe that, on average, scores are indeed slightly higher for performances by skaters with a compatriot judge on the panel the by those without. Second, despite some fluctuations in the order of magnitude that is statistically to be expected, JGP and Non-JGP events do seem to follow roughly similar pre-trends in the three seasons before the reform in our data, thus corroborating our difference-

in-difference identification strategy. Third, there is no indication for a decreasing compatriot score advantage in treated events (Non-JGP) following the transparency reform compared to non-treated events (JGP).

Table 4.A.3 presents our formal regression results that implement the estimation strategy described in equation 4.5.2. For both artistic score and technical score, we estimate a statistically significant baseline compatriot advantage even when including skater fixed effects, which account for the fact that more judges come from countries like Russia that are traditionally strong in figure skating. We estimate a baseline bias of about 0.08-0.09 within-round SDs for both the artistic score and the technical score in favor of skaters with a compatriot judge on the panel. Further including control variables for the skater's current world rank at the time of performance reduces the estimated bias to around 0.06 within-round SDs. There seems to be little difference between JGP and Non-JGP events pre-reform, despite the fact that individual judges' scores from JGP events already being published openly prior to the 16/17 season.

Importantly, we find no evidence for a decrease in the average compatriot bias for treated Non-JGP events relative to JGP events after the reform in 2016. The estimated coefficient of 0.026 for the artistic score is statistically insignificant and goes in the wrong direction. Importantly, the implied estimate for the post-reform compatriot bias at Non-JGP events is positive (0.057) and remains statistically different from zero ($p = 0.001$) based on the coefficients in column (2). For the technical score, the point estimate is even large and positive, although far from statistically significant. Thus, we conclude that the transparency reform seems to have been unsuccessful in achieving one of its main objectives, i.e. to reduce nationalistic favoritism in figure skating.

This null result is consistent with our theoretical model from section 4.3, which predicts that a reduction in individual judges' favoritism may be offset in the aggregate score by conformity motives of other judges. Another explanation could be that transparency triggers opposing motives for judges' evaluations. For example, public scrutiny and fairness norms would push biased judges to curb their tendencies for favoritism, whereas audiences in the home country as well as national associations that appoint the judges may in fact expect that judges behave in a biased way by skewing scores for their compatriot skaters upwards. Finally, it is also possible that judges become more strategic in the way they award scores such that they become less biased for performances in which their score is not pivotal and more biased for performances that actually matter for the final ranking of the skaters.

4.7.2 Heterogeneous effects

While we find no evidence for a decrease in the compatriot score advantage on average following the transparency reform, it is conceivable that publishing individual judges' scores has different effects on nationalistic judging depending on character-

Table 4.7.1. Effect of the transparency reform on compatriot score advantage

	Artistic score		Technical score	
	(1)	(2)	(3)	(4)
Compatriot	0.089***	0.060***	0.083***	0.059**
	(0.021)	(0.023)	(0.025)	(0.027)
Compatriot × Non-JGP	-0.012	0.013	-0.033	-0.010
	(0.029)	(0.031)	(0.037)	(0.038)
Compatriot × Post	-0.053*	-0.043	-0.072**	-0.066**
	(0.027)	(0.028)	(0.033)	(0.033)
Compatriot × Post × Non-JGP	0.046	0.030	0.064	0.054
	(0.039)	(0.040)	(0.050)	(0.050)
Home event		0.104***		0.138***
		(0.020)		(0.024)
Controls for current world rank	—	Yes	—	Yes
Skater FEs	Yes	Yes	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16764	16764	16764
R^2	0.867	0.875	0.708	0.711

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

istics of the competition, such as how prestigious the event is and how much public attention it thus generates.

To check for potential heterogeneity along this dimension, we construct a proxy for the prestigiousness of a round using the average world rank of all participating skaters at the time of the competition. We then include interactions with this variable to examine whether the compatriot advantage in rounds with higher average quality of skaters decreases more strongly in response to the reform. Appendix Table 4.A.2 shows that this does not seem to be the case. While the estimated coefficients for $Roundquality \times Compatriot \times Non - JGP \times Post$ is negative as predicted, it is quantitatively small and statistically insignificant in all specifications.

4.8 Additional Results

4.8.1 Consistency of scores by individual judges

The artistic score awarded by judges are calculated from several sub-scores for different components of the performance, e.g. skating skills, transitions, interpretation of music. Likewise, the technical score is calculated from grades of execution for each technical element (e.g jump, spin) performed by the skater. When all scores

are made public in a deanonymized way, judges may also try to signal their competence by reporting more similar evaluations for the different score component, as discussed in section 4.3.3.

We evaluate this empirically by using performance-judge-level score data and computing the within-judge standard deviation across all components of the artistic score as outcome variable. For the technical score, we compute the within-judge standard deviation across all GOEs of a performance. Using the same difference-in-differences approach as before, we find that indeed judges become more consistent in their evaluations for artistic score components. Columns (1) and (2) in Appendix Table show that after the transparency reform, the standard deviation of artistic components drops by 0.014, or 6% of the baseline, for Non-JGP performances compared to JGP performances. This effect is statistically significant at the 1% level. However, we find no effect of transparency on within-judge consistency of grades awarded for the different technical elements. Hence, our results on within-judge consistency are analogous to the previous findings on the score dispersion across judges in a panel, in that we only find effects for the more subjective and more deliberately assigned artistic scores, but not for the more objective and more spontaneously assigned technical scores.

It remains somewhat ambiguous whether more consistent scores are an indicator for more or less accurate performance evaluations by judges. On the one hand, very similar scores for each component may indeed mean less arbitrariness and more confidence in judging behavior. On the other hand, it may also simply result from "lazy" judging, in particular if a judge settles on one score for all components. Tellingly, one exception to the general decrease in dispersion across artistic score components is that the share of cases in which a judge simply assigns the same score for each component drops by more than 50% in Non-JGP events following the reform.[25] Thus, our results support social image concerns as an important mechanism through which transparency affects judge behavior, but the increase in consistency would not be predicted by other potential mechanisms such as collusion or strategic (counter-)exaggerations behavior.

4.8.2 Social Learning and Effort Exertion

Given that judges are not allowed to communicate with each other throughout the round, one might wonder how judges can award more similar scores post-reform. Two main possibilities are social learning and effort exertion.

Because the overall scores awarded to any particular skater are shown on on-site screens after the performance (during the performance for the technical score),

25. While the incidence of such "straightlining" cases was 0.41% for Non-JGP events in the 2012/13 to 2015/16 seasons, it drops to 0.19% from season 2016/17 onwards. For JGP events, the incidence was 0.14% pre-reform and 0.11% post-reform.

judges can improve their impression of how other judges on the panel score performances as the round advances. Thus, under social learning, performances later in the round would receive more similar scores. If the conformity effect is mainly driven by social learning, we might expect the effect of the reform to be more pronounced for performances later in the round. In contrast, an effort-exertion-based mechanism does not yield such clear-cut predictions. For instance, it might be easier to exert effort earlier in the round, when judges are not yet fatigued, so that the reform leads to a stronger decrease in standard deviation for performances earlier in the round. On the other hand, it is also possible that judges require some time to settle into their roles, so that it is easier to exert effort for later performances. Nonetheless, if the reform is driven by social learning, we should not expect to see standard deviation decrease more for earlier performances.

In this section, we thus see if, and how the effect of reform varies with within-round starting order by interacting the standard DiD specification with starting number. However, starting numbers are not completely randomly assigned– better skaters are usually scheduled to skate later. Typically, skaters are placed into groups (starting-order groups) based on their world rank or placement in the short program, so that higher-ranked/placed skaters are assigned to later starting-order groups. Skaters then draw for starting numbers within these groups, so that within-group, skater ability is uncorrelated with starting order. [26][27] We thus include *Round* $\times$ *Group* FEs and estimate the following triple difference-in-differences specification:

$$\begin{aligned} y_{irdp} \;=\; & \alpha_i + \beta_1 \cdot StNr_{irdp} + \beta_2 \cdot StNr \times NonJGP_{irdp} \\ & + \beta_3 \cdot StNr \times Post_{irdp} + \beta_4 \cdot StNr \times NonJGP \times Post_{irdp} \\ & + \beta_5 \cdot score_{irdp} + \delta_{rd} + \varepsilon_{irdp}\,, \end{aligned} \tag{4.8.1}$$

where $StNr_{irdp}$ is skater i's starting number in round r and starting-order group d, and δ_{rd} is round $\times$ starting-order group FEs. Because starting order has been shown to influence judges' scores, we also control for the relevant performance score. [28] If the conformity effect in Section 4.6 is caused by social learning, we should expect β_4 to be negative, so that post-reform, decrease in standard deviation within starting order groups is steeper for Non-JGP rounds, relative to JGP rounds.

26. Pooling short- and long-program rounds, starting-order groups tend to be larger for JGP rounds (14), compared to Non-JGP rounds (6.5). This is because JGP short program rounds have completely randomized starting numbers. Draw group sizes are similar for the long program (3.9 for both JGP and Non-JGP rounds).

27. Because Grand Prix Series and Final events often use the reverse order of world rankings or short order placement to determine the skating order, we exclude these events from the analysis.

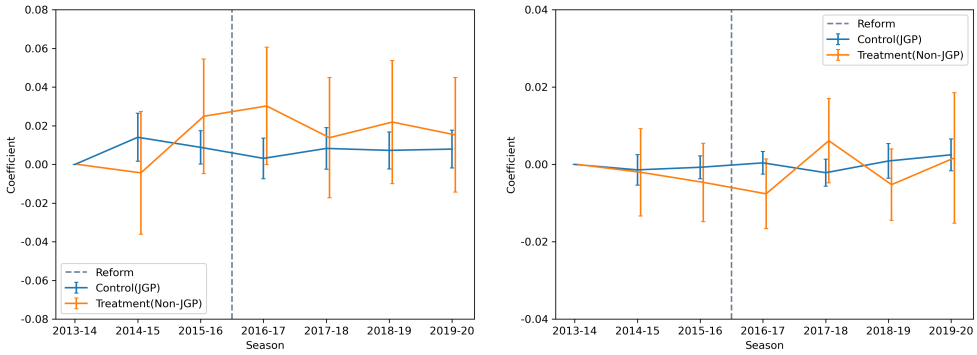
28. Looking at rounds with randomized starting numbers, Bruine de Bruin (2006) finds that later skaters tend to obtain higher scores. Note that Bruine de Bruin (2006)'s study uses rounds from 1994 to 1999, where the judging system used involves ranking skaters within the round. (6.0 Judging System)

Table 4.8.1. Heterogeneous effects by starting number

	SD of Artistic Score		SD of Technical Score	
	(1)	(2)	(3)	(4)
StNr	0.002	0.000	0.000	0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Post × StNr	-0.003	-0.002	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Non-JGP × StNr	-0.020***	-0.015***	-0.002	-0.000
	(0.005)	(0.005)	(0.003)	(0.004)
Non-JGP × Post × StNr	0.020***	0.015**	0.005	0.002
	(0.007)	(0.007)	(0.004)	(0.005)
Artistic Score	0.004***	-0.015***		
	(0.001)	(0.003)		
Technical Score			0.027***	0.025***
			(0.000)	(0.001)
Constant	1.716***	2.365***	0.260***	0.296***
	(0.052)	(0.106)	(0.023)	(0.032)
Skater FEs	—	Yes	—	Yes
Observations	12858	12784	12858	12784
R^2	0.401	0.550	0.739	0.787

Coefficients are obtained from estimating equation 4.8.1, with standard deviation of panel scores as dependent variable. Standard errors in parentheses are clustered at event level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

From column (1) of Table 4.8.1, we see that the standard deviation increases (insignificantly) with starting number for JGP rounds pre-reform. For Non-JGP rounds, the standard deviation decreases (significantly) as the round proceeds, so that judges tend to award more similar scores towards skaters skating later in the starting order group. Scaling by the average number of skaters in a starting order group, this would imply a decrease of 0.078 from the first to the last skater in the group. This could be due to social learning, as judges acquire panel-specific information on scoring with each additional skater, or effort exertion, if judges need some time to acclimate themselves to judging and subsequently put more effort into judging later performances, or a combination of both of these factors. This scoring pattern might be absent from JGP competitions because public attention in these competitions are lower, so that judges feel less pressure to award a consensus score. The estimate on $Non - JGP \times Post \times StNr$ is positive, and similar in absolute value to the estimated coefficient on $Non - JGP \times StNr$ suggesting that the tendency to award more similar scores towards later performances in Non-JGP rounds disappears post-reform. Including skater FEs in col (2) does not change the estimates much. Columns (3)



(a) SD of artistic score **(b)** SD of technical score

Figure 4.8.1. Effect of StNr on panel standard deviation for JGP (Control) and Non-JGP (Treat) events from seasons 2013-14 to 2019-20.

Notes: We regress Panel standard deviation on StNr, interacted with season, controlling for the score and including *Round* × *Group* FEs. Standard errors are clustered at the event level. The blue (orange) points are the coefficients on StNr for JGP/Control (Non-JGP/Treatment) performances. The dashed line indicates the implementation of the transparency reform; error bars are 95% confidence intervals.

and (4) estimate 4.8.1 with the standard deviation of the technical score as the dependent variable. Though insignificant, the estimated coefficients share the same signs as those in columns (1) and (2).

Though this would appear to hint at an effort-exertion-based mechanism, we are hesitant to draw any conclusions. From Figure 4.8.1, the increase in the estimated coefficient on *StNr* for Non-JGP rounds seems to occur in the season preceding the reform. Furthermore, within each season, the estimated coefficients on *StNr* do not differ significantly between JGP and Non-JGP rounds. Nonetheless, we can conclude that, for skaters of ex-ante equal ability, the conformity effect does not seem to vary with starting number, which points us to other explanations. Firstly, it does not seem to be the case that judges try to emulate other judges on the panel in order to award more similar scores. Rather, it seems that judges likely already have a common consensus score towards which they move post-reform. At the same time, higher effort or attention remains another explanation, as it is possible that post-reform, judges exert greater effort in grading all performances in the round, which would lead to a uniformly lower within-in panel standard deviation.

4.8.3 Changes in Selection of Judges

The selection of judges to serve in a panel is not completely random. For JGP (Control) and some of the Non-JGP (Treat) events,[29] judges are selected by the organizing country, subject to several restrictions. Organizing countries are required to

29. Notably, judges from the Grand Prix Series are not randomly selected. However, the small size of these events means that they only account for a fraction of the sample. Results are robust to dropping these observations from the Treatment group.

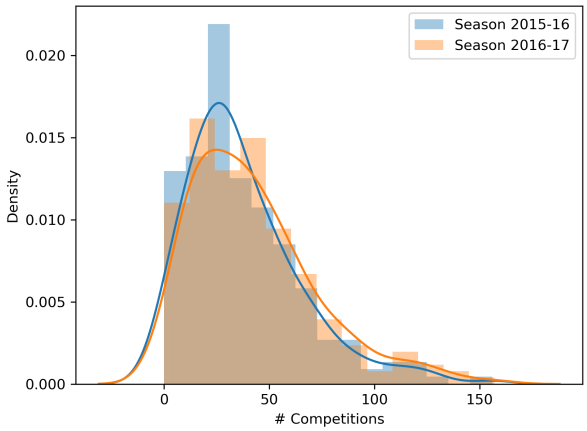


Figure 4.8.2. Distributions of Non-JGP judge experience pre- and post-reform

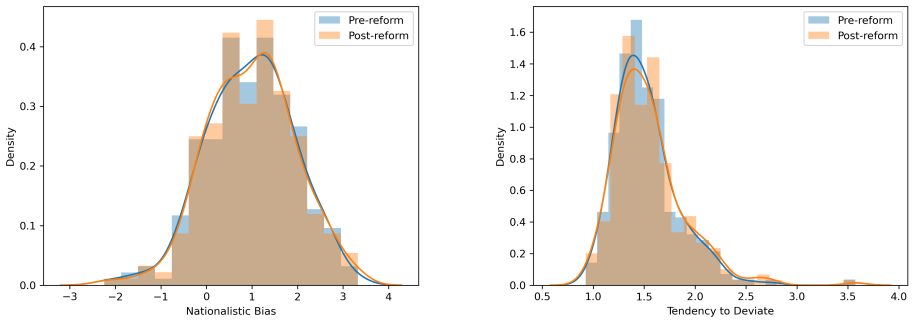
Notes: Judge experience in a season is proxied by the number of competitions he or she has judged at, from season 2005-06 up to the current season.

select judges from a pool of qualified individuals ('International Judges') and are not allowed to have more than one judge from their country serving in a given competition. For the remaining Non-JGP (Treat) events, judges are selected in a two-step procedure— firstly, each national skating federation nominates a judge from their country to serve in a particular competition (e.g. World Figure Skating Championships Womens); next, the ISU randomly draws from the pool of proposed candidates the panel of judges that will serve in a particular competition.

Therefore, the observed decrease in score dispersion could be caused by changes in the selection criteria of organizing countries (JGP and GP Series) or national skating federations (Non-GP Treatment events; i.e. all other events)— for instance, because these countries or federations might feel compelled to select or propose judges that are more impartial or experienced when scores are not anonymous, or because some judges become less willing to serve in panels. We provide several pieces of evidence that speak against this selection mechanism.

Firstly, for the subset of Non-JGP (Treat) events where national federations are supposed to submit a judge from their country to the random draw (Non-GP Treat), we find no major changes in the pool of countries submitting candidates to Non-GP Treat events. Table 4.A.4 shows that, overall, approximately 80% of countries who submitted a judge in a particular season will submit a judge in the next season, and this figure remains constant in the post-reform seasons. Furthermore, looking at the decomposition of these transitions by the type of event countries submit to (e.g. Four Continents, World Championships, etc.), there also appears to be no major in- or outflux of countries after the reform.

To check if countries become more likely to select more experienced judges, we use the number of competitions since the 2005-06 season (the earliest season in the



(a) Distribution of Judge Nationalistic Bias **(b)** Distribution of Judge Conformity

Figure 4.8.3. Distributions of Non-JGP judge characteristics pre- and post-reform

Notes: Nationalistic bias is defined as the difference in the average deviation from other judges' scores on the panel when the skater is compatriot and when skater is not compatriot; tendency to deviate is defined as the overall tendency of the judge to deviate from the scores given by other judges on the panel; both measures are based on JGP (Control) data, from seasons 2005-06 to 2012-13.

dataset) in which a judge has served in a panel to construct a proxy for experience by judge and season. Figure 4.8.2 compares histograms of judge experience in the 2015-16 and the 2016-17 seasons. We do not observe strong evidence that the distribution shifts significantly from pre-reform to post-reform (p-value of Kolmogorov-Smirnov test = 0.1397).[30]

Next, we investigate whether more impartial judges are selected after the transparency reform, either by organizing countries (in GP events) or national skating federations (all other treated events). As measure of (nationalistic) judge impartiality, we calculate for each judge the difference in deviation from other judges' scores when the skater is a compatriot compared to when the skater is not a compatriot, over the 2005-06 season to the 2012-13 season. Because the pre-reform Non-JGP results are anonymized, we can compute this measure using only data from the JGP (Control) results.[31] Although not all judges in our sample have served in a JGP event, we could obtain this measure for approximately 80% of judges. Figure 4.8.3 (a) plots the distributions of this impartiality measure for judges in Non-JGP events in the last pre-reform season and the first post-reform season.[32]

Lastly, it is possible that judges who are less conformist might opt out of judging in Non-JGP events post-reform, knowing that their actions are now more scrutinized by the public. We use as a measure of conformity the overall tendency of the judge to

30. See Appendix Figure 4.A.4 for histograms of judge experience in all seasons.

31. Note that under pre-reform anonymization, information regarding judges' impartiality and conformity in Non-JGP competitions are also concealed from national skating federations and organizing countries, so that judges' behavior in JGP competitions is the only credible source of information for these federations to decide which judges to select.

32. See Appendix Figure 4.A.6 for histograms of judge impartiality in all seasons.

deviate from the scores given by his/her fellow judges. As with the impartiality measure, coverage is around 80%, and there does not appear to be a significant shift in the overall distribution of conformity from the pre-reform period to the post-reform period.[33] Alternatively, it could also be the case that more reputation-concerned judges become less likely to opt into of Non-JGP events post-reform. While we cannot measure judges' reputation consciousness, Based on the subsample of judges who have served in at least one pre-reform Non-JGP event, Appendix Tables 4.A.5 and 4.A.6 show that there is no drastic extensive or intensive margin decrease in judges' propensity to serve in Non-JGP event after the transparency reform.

4.9 Conclusion

In this paper, we studied the effect of transparency on performance evaluation in committees in a high-stakes, professional context. Specifically, we evaluated a reform implemented in the sport of figure skating that increased the visibility of judges' decisions. Prior to the reform, judges' scores were published anonymously, thus shielding the judge from public censure or supervision. While this prevents judges from being swayed by public opinion and coerced into collusion by their fellow judges, this opacity also made it was relatively easy for judges to engage in nationalistic favoritism, so that, following accusations of nationalistic judging in the 2014 Sochi Olympics, the ISU de-anonymized result publication for all events.

To illustrate how increased visibility might impact judges' scoring behavior, we proposed a theoretical framework à la Morris and Shin (2002) with potentially biased and conformist judges, in which the transparency reform enters as an increase in conformist concerns. In line with the predictions of the model, we find that the within-performance score dispersion for artistic scores decreases sharply post-reform, indicating that judges tend to award more similar scores. In further support of a conformity-based explanation, we also see that this effect is stronger in settings with greater public attention, where judges might feel higher pressure to conform. Lastly, we find that skaters are scored higher when they have a compatriot judge on the panel, and that this compatriot advantage does not decrease post-reform. This is, at first glance, perhaps surprising, given that the reform was implemented precisely to address such concerns. However, this finding is compatible with our model's predictions, and highlights the limited impact that greater transparency can have on aggregate biases in committee decisions.

Though the sharp increase in scoring similarity is in line with previous research in different contexts, the inability of judges to communicate with each other in our setting rules out informational exchange or persuasion as mechanisms driving the conformity effect we see. Similarly, we do not find any evidence of social learning

33. See Appendix Figure 4.A.5 for histograms of judge conformity in all seasons.

in our setting. Our model instead suggests two potential sources for this result—increased effort leading to higher signal precision, or herding on a common prior—with largely different welfare consequences. The former leads to less arbitrary and random scoring, whereas the latter has the opposite effect, and could over time lead to a more entrenched system where performances by rookie skaters are insufficiently rewarded. We ultimately cannot distinguish between these channels with our data, and leave this as a potential avenue to explore in future research.

In general, transparency, by activating social image concerns, is a powerful tool that can be used to align individual behavior with public norms and expectations. Whether this can be successfully utilized to achieve desirable committee outcomes, however, likely depends on a variety of factors. These include, among others, the prevailing norms in the society, the degree of subjectivity of the decision, and the composition of the committee, which influence the quality of decisions made under transparency. Thus, policy makers should carefully consider the context when implementing transparency policies. However, one advantage of higher transparency is hardly disputable: it generates publicly available data for third parties like journalists and researchers and thereby potentially long-term value.

Appendix 4.A Supplementary Figures and Tables

ISU European Championships 2014

MEN FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	Javier FERNANDEZ	ESP	20	175.55	88.19	87.36	0.00

#	Executed Elements	Info	Base Value		GOE	The Judges Panel (in random order)									Ref	Scores of Panel
1	4T		10.30		-0.43	-1	0	-1	-1	1	-2	0	2	-1		9.87
2	4S+3T<	<	13.40		-0.43	0	0	-1	-1	0	-1	0	1	-1		12.97
3	3A		8.50		1.71	1	1	2	0	2	1	3	3	2		10.21
4	CSSp4		3.00		0.57	1	1	1	1	1	1	2	2	1		3.57
5	StSq3		3.30		0.79	2	1	2	1	1	2	2	2	1		4.09
6	4S		11.55	x	-2.00	-2	-2	-2	-2	-2	-2	-2	-1	-2		9.55
7	2Lz+2T		3.74	x	0.04	0	0	0	0	0	0	1	1	0		3.78
8	3Lo		5.61	x	0.80	1	1	2	1	1	1	1	2	1		6.41
9	3F+1Lo+3S		11.00	x	0.50	1	1	0	0	1	0	2	2	0		11.50
10	FCCoSp4		3.50		0.29	1	0	0	0	1	1	1	1	0		3.79
11	ChSq1		2.00		1.50	2	2	3	2	2	2	3	2	2		3.50
12	3S		4.62	x	0.40	1	0	0	0	1	0	1	2	1		5.02
13	CCoSp4		3.50		0.43	1	1	1	1	1	0	1	1	0		3.93
			84.02													**88.19**
	Program Components				Factor											
	Skating Skills				2.00	8.75	8.75	8.50	8.25	8.25	8.50	9.25	8.75	7.75		8.54
	Transition / Linking Footwork				2.00	9.50	8.75	8.75	8.25	8.75	8.75	8.75	8.00	8.00		8.57
	Performance / Execution				2.00	9.00	9.00	9.00	8.50	9.00	8.50	9.00	8.50	9.00		8.86
	Choreography / Composition				2.00	8.75	9.00	9.00	8.50	8.50	8.75	9.50	8.25	8.50		8.71
	Interpretation				2.00	9.50	9.25	9.25	8.50	9.00	9.00	9.50	8.50	8.25		9.00
	Judges Total Program Component Score (factored)															**87.36**
	Deductions:															**0.00**

< Under-rotated jump x Credit for highlight distribution, base value multiplied by 1.1

(a) Pre-reform

ISU European Figure Skating Championships 2017

MEN FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	**Javier FERNANDEZ**	**ESP**	**22**	**190.59**	**98.29**	**93.30**	**1.00**

#	Executed Elements	Info	Base Value		GOE	J1	J2	J3	J4	J5	J6	J7	J8	J9	Ref	Scores of Panel
1	4T		10.30		2.71	2	3	2	2	3	3	3	3	3		13.01
2	4S+2T		11.80		-0.20	-2	1	0	0	0	0	1	-1	-1		11.60
3	3A+3T		12.80		0.86	1	1	1	1	0	1	2	-1	1		13.66
4	CSSp3		2.60		0.43	1	1	1	1	0	1	2	0	1		3.03
5	ChSq1		2.00		1.50	2	2	1	2	2	3	3	2	2		3.50
6	4S		11.55	x	-4.00	-3	-3	-2	-3	-3	-3	-3	-3	-3		7.55
7	3A		9.35	x	-0.86	-2	0	-1	-1	-1	-1	2	-1	-1		8.49
8	3Lz		6.60	x	1.10	1	2	1	1	2	1	3	2	2		7.70
9	3F+1Lo+3S	!	11.22	x	0.00	0	0	0	0	-1	0	1	0	0		11.22
10	FCCoSp4		3.50		0.36	1	1	0	1	0	1	1	1	0		3.86
11	3Lo		5.61	x	-0.80	-2	-1	-1	-1	-2	-1	-1	-1	-1		4.81
12	StSq4		3.90		1.60	3	2	1	2	2	3	3	2	2		5.50
13	CCoSp4		3.50		0.86	2	1	1	2	1	2	2	2	2		4.36
			94.73													**98.29**
	Program Components				Factor											
	Skating Skills				2.00	9.50	9.25	8.75	9.25	9.00	9.00	9.50	9.50	9.50		9.29
	Transitions				2.00	9.75	9.00	8.75	9.25	8.75	9.00	9.50	9.25	9.50		9.18
	Performance				2.00	9.75	9.00	9.00	9.50	9.00	8.00	9.50	9.25	9.25		9.21
	Composition				2.00	9.50	9.50	9.25	9.50	9.25	9.25	10.00	9.50	9.50		9.43
	Interpretation of the Music				2.00	10.00	9.25	8.50	9.50	9.50	9.50	10.00	9.50	9.50		9.54
	Judges Total Program Component Score (factored)															**93.30**
	Deductions				Falls: -1.00 (1)											**-1.00**

x Credit for highlight distribution, base value multiplied by 1.1 ! Not clear edge

(b) Post-reform

Figure 4.A.1. Online publication of results for Non-JGP (Treat) events pre- and post-reform.

Notes: Notice that the order of panel judges is not revealed in panel (a), while it is revealed in panel (b). This order can be linked back to the individual judges on the panel.

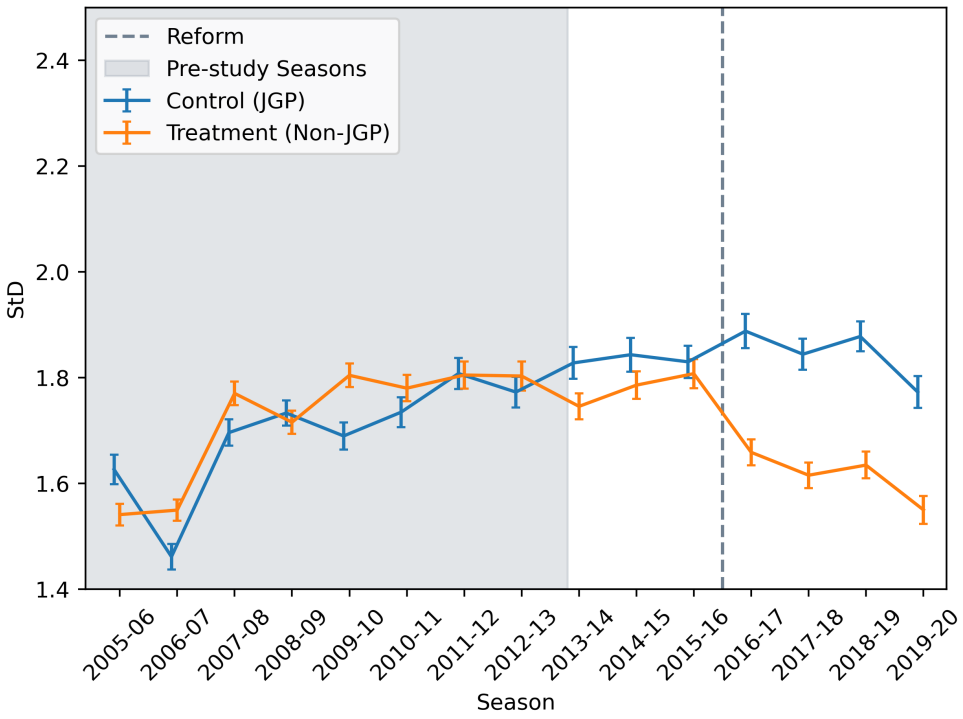
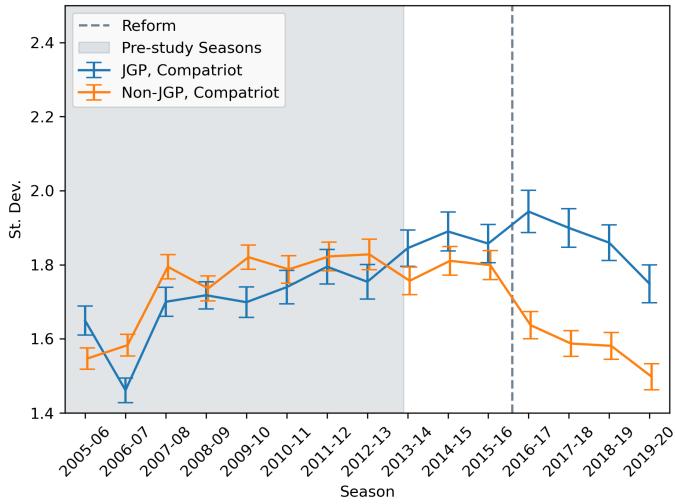
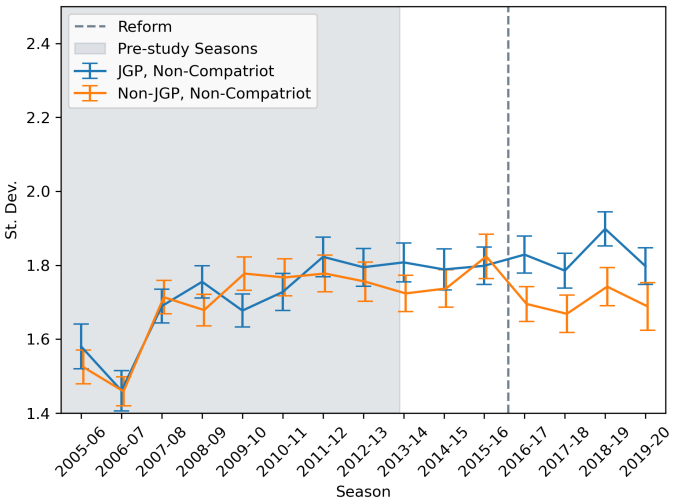


Figure 4.A.2. Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2005-06 to 2019-20

Notes: Each orange(blue) point plots the average panel standard devation for treatment(control) performances in a season, over the seasons 2005-06 to 2019-20. The dashed line indicates implementation of the transparency reform, from the 2016-17 season onwards.



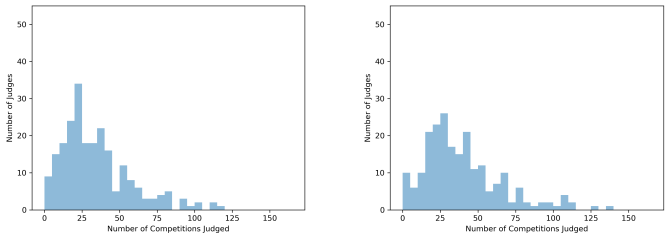
(a) Compatriot



(b) Non Compatriot

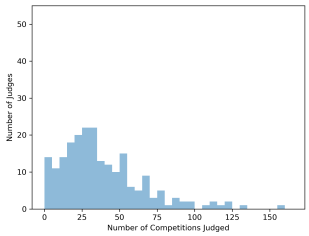
Figure 4.A.3. Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2013-14 to 2019-20, split by presence of compatriot judge on panel.

Notes: Each orange(blue) point plots the average panel standard deviation for treatment(control) performances in a season, over the seasons 2005-06 to 2019-20. The dashed line indicates implementation of the transparency reform, from the 2016-17 season onwards.

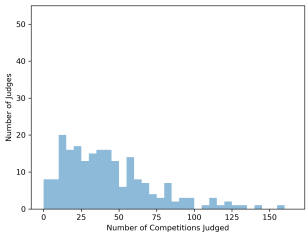


(a) Season 2013-14

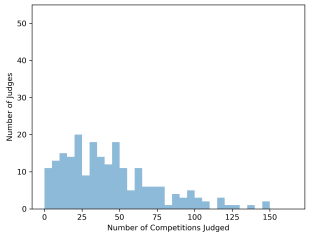
(b) Season 2014-15



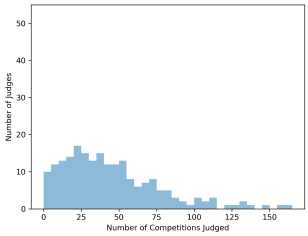
(c) Season 2015-16



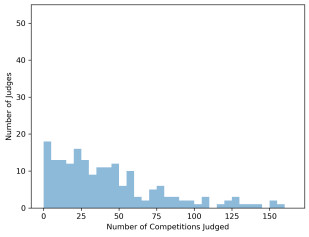
(d) Season 2016-17



(e) Season 2017-18



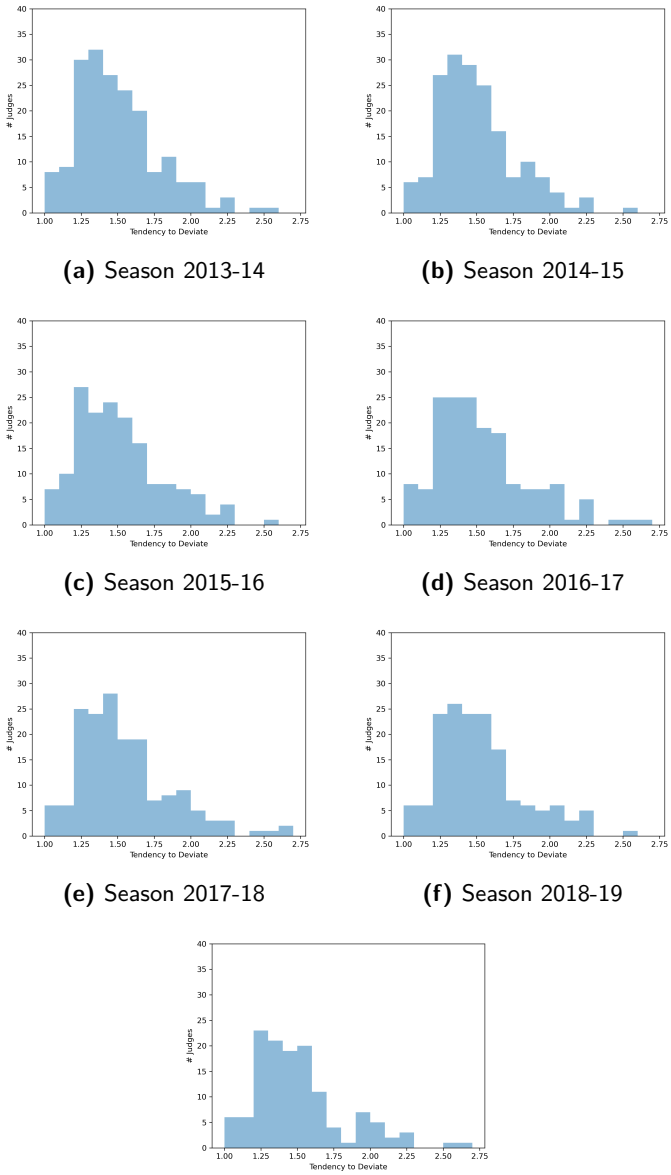
(f) Season 2018-19



(g) Season 2019-20

Figure 4.A.4. Distributions of Non-JGP (Treat) judge experience by season, from seasons 2013-14 to 2019-20.

Notes: Judge experience in a season is computed as the number of competitions he/she has judged at up until that season.



(a) Season 2013-14 **(b)** Season 2014-15

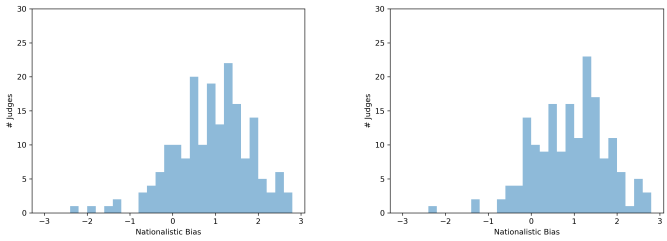
(c) Season 2015-16 **(d)** Season 2016-17

(e) Season 2017-18 **(f)** Season 2018-19

(g) Season 2019-20

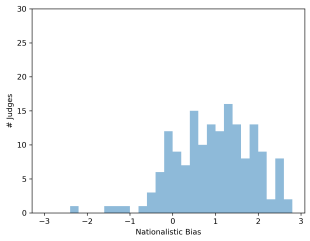
Figure 4.A.5. Distributions of Non-JGP (Treat) judge conformity by season, from seasons 2013-14 to 2019-20.

Notes: For each judge, his/her measure of deviation is the average deviation of all performances where he/she has judged in, where his/her deviation in a performance is calculated as the absolute value of his score from that of the leave-one-out panel mean.

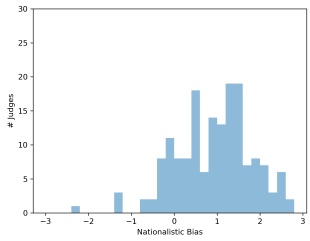


(a) Season 2013-14

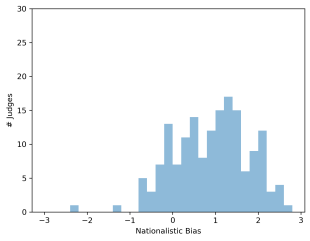
(b) Season 2014-15



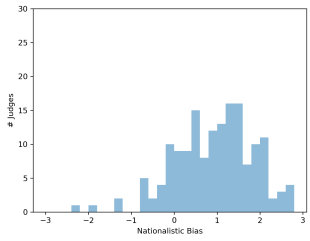
(c) Season 2015-16



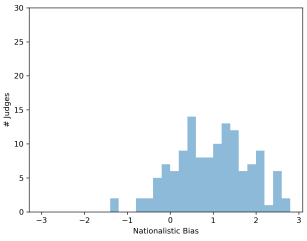
(d) Season 2016-17



(e) Season 2017-18



(f) Season 2018-19



(g) Season 2019-20

Figure 4.A.6. Distributions of Non-JGP judge nationalistic partiality by season, from seasons 2013-14 to 2019-20.

Notes: For each judge, his/her measure of (nationalistic) impartiality is the average deviation from the leave-one-out panel mean when the skater is compatriot, minus the the average deviation from the leave-one-out panel mean when the skater is non-compatriot.

Table 4.A.1. Effect of de-anonymized publication on standard deviation of panel scores.

	(1)	(2)	(3)
Intercept	1.7254***	1.7265***	1.7346***
	(0.0130)	(0.0130)	(0.0150)
Post	0.1192***	0.1192***	0.1198***
	(0.0260)	(0.0260)	(0.0260)
Non-JGP	0.0046	0.0050	0.0041
	(0.0170)	(0.0170)	(0.0170)
Post × Non-JGP	-0.2325***	-0.2325***	-0.2333***
	(0.0330)	(0.0330)	(0.0330)
Skater FE	—	—	—
Discipline × Segment	—	—	—
Skater Home Event	—	Yes	Yes
Starting Order	—	—	Yes
Observations	38,677	38,677	38,677
R^2	0.0098	0.0098	0.0098
Mean	1.7300	1.7300	1.7300

Notes. Estimates of 4.5.1, with standard deviation of panel scores as dependent variable, using performances from seasons 2005-06 to 2019-20. Standard errors clustered at round (e.g. Olympics 2018 Women's Free Skate) level. Mean refers to the pre-reform average panel standard deviation for Non-JGP performances.

Table 4.A.2. Effect of the transparency reform on compatriot score advantage

	Artistic score		Technical score	
	(1)	(2)	(3)	(4)
Compatriot	0.089***	0.060***	0.083***	0.059**
	(0.021)	(0.023)	(0.025)	(0.027)
Compatriot × Non-JGP	-0.002	0.020	-0.028	-0.007
	(0.037)	(0.038)	(0.042)	(0.042)
Compatriot × Post	-0.053*	-0.043	-0.072**	-0.066**
	(0.027)	(0.028)	(0.033)	(0.033)
Compatriot × Post × Non-JGP	0.040	0.028	0.054	0.046
	(0.046)	(0.046)	(0.057)	(0.057)
Round quality × Comp. × Non-JGP	0.014	0.009	0.007	0.004
	(0.014)	(0.015)	(0.015)	(0.015)
Round quality × Comp. × Non-JGP × Post	-0.009	-0.004	-0.013	-0.010
	(0.017)	(0.017)	(0.021)	(0.020)
Home event		0.104***		0.138***
		(0.020)		(0.024)
Controls for current world rank	—	Yes	—	Yes
Skater FEs	Yes	Yes	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16764	16764	16764
R^2	0.867	0.875	0.708	0.711

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.A.3. Effect of the transparency reform sub-score consistency

	Artistic score components		Technical score components		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.004	0.004	-0.003	-0.021	-0.025
	(0.004)	(0.004)	(0.012)	(0.013)	(0.015)
Post × Non-JGP	-0.013***	-0.014***	-0.014	-0.018	0.006
	(0.005)	(0.005)	(0.016)	(0.016)	(0.017)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	169242	167136	169214	167109	108675
R^2	0.700	0.733	0.308	0.415	0.337

Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.A.4. Statistics on pool of countries submitting judges to Non-GP treatment events.

Event Type	# Country	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20
European Championships	Outgoing	3	3	1	4	3	3	N.A.
	From Previous Season	N.A.	23	24	25	25	24	24
	Incoming	N.A.	4	2	4	2	3	4
	Total	26	27	26	29	27	27	28
Four Continents	Outgoing	7	11	8	7	8	9	N.A.
	From Previous Season	N.A.	20	19	20	19	20	20
	Incoming	N.A.	10	9	6	9	9	6
	Total	27	30	28	26	28	29	26
World Juniors	Outgoing	7	5	7	5	5	7	N.A.
	From Previous Season	N.A.	23	25	24	25	27	23
	Incoming	N.A.	7	6	6	7	3	6
	Total	30	30	31	30	32	30	29
World Championships	Outgoing	4	5	5	3	6	9	N.A.
	From Previous Season	N.A.	25	23	21	23	23	21
	Incoming	N.A.	3	3	5	6	7	8
	Total	29	28	26	26	29	30	29
Total	Outgoing	21	24	21	19	22	28	N.A.
	From Previous Season	N.A.	91	91	90	92	94	88
	Incoming	N.A.	24	20	21	24	22	24
	Total	112	115	111	111	116	116	112

Table 4.A.5. Proportion of Non-JGP (Treatment) judges remaining next season.

Season	# Judges	% Remaining Next Season	Difference Next Season	T-test p-value
2005-06	245	0.706	0.046	0.257
2006-07	238	0.752	-0.11	0.009
2007-08	240	0.642	0.054	0.228
2008-09	207	0.696	-0.052	0.248
2009-10	230	0.643	0.019	0.682
2010-11	216	0.662	0.044	0.332
2011-12	214	0.706	0.056	0.189
2012-13	222	0.761	-0.045	0.277
2013-14	229	0.716	-0.069	0.116
2014-15	218	0.647	0.028	0.545
2015-16	215	0.674	0.049	0.268
2016-17	210	0.724	-0.085	0.06
2017-18	216	0.639	-0.048	0.316
2018-19	208	0.591	N.A.	N.A.

Table 4.A.6. # Competitions by Non-JGP (Treatment) judges Who remain in next season.

Season	# Competitions Season	# Competitions Season + 1	Difference	T-test p-value
2005-06	5.734	4.965	-0.769	0.057
2006-07	5.067	5.017	-0.050	0.889
2007-08	5.286	5.143	-0.143	0.731
2008-09	5.118	5.201	0.083	0.853
2009-10	5.297	4.642	-0.655	0.124
2010-11	4.937	5.238	0.301	0.465
2011-12	5.060	4.589	-0.470	0.245
2012-13	4.219	4.941	0.722	0.085
2013-14	4.817	4.207	-0.610	0.152
2014-15	4.482	4.447	-0.035	0.935
2015-16	4.566	4.821	0.255	0.549
2016-17	4.724	5.493	0.770	0.110
2017-18	4.775	4.638	-0.138	0.774
2018-19	4.821	4.545	-0.276	0.566

Table 4.A.7. Impartiality Coverage

Total	Not Found	Found	Percent Found
229	40	189	0.825328
219	44	175	0.799087
215	49	166	0.772093
210	41	169	0.804762
218	50	168	0.770642
208	45	163	0.783654
189	58	131	0.693122

Table 4.A.8. Conformity Coverage

Total	Not Found	Found	Percent Found
229	39	190	0.829694
219	44	175	0.799087
215	49	166	0.772093
210	41	169	0.804762
218	50	168	0.770642
208	45	163	0.783654
189	58	131	0.693122

References

Asch, Solomon E. 1951. "Effects of group pressure upon the modification and distortion of judgment." In *Groups, leadership and men; research in human relations*. Edited by H. Guetzkow. Pittsburgh: Carnegie Press, 177–190. [162]

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review* 107 (4): 1207–1238. [166]

Bagues, Manuel F., and Berta Esteve-Volart. 2010. "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment." *Review of Economic Studies* 77 (4): 1301–1328. [166, 167]

Baker, Michael, Yosh Halberstam, Kory Kroft, Alexandre Mas, and Derek Messacar. 2019. "Pay Transparency and the Gender Gap: Working Paper." [166]

Bar-Isaac, Heski. 2012. "Transparency, Career Concerns, and Incentives for Acquiring Expertise." *The B.E. Journal of Theoretical Economics* 12 (1): [163]

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015. "Unintended Effects of Anonymous Résumés." *American Economic Journal: Applied Economics* 7 (3): 1–27. [166]

Benesch, Christine, Monika Bütler, and Katharina E. Hofer. 2018. "Transparency in parliamentary voting." *Journal of Public Economics* 163: 60–76. [162, 165, 176]

Bertrand, Marianne, Sandra E. Black, Sissel Jensen, and Adriana Lleras-Muney. 2018. "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway." *Review of Economic Studies* 86 (1): 191–239. [166]

Böheim, René, Mario Lackner, and Wilhelm Wagner. 2020. "Raising the Bar: Causal Evidence on Gender Differences in Risk-Taking from a Natural Experiment." *IZA Discussion Paper No. 12946*, [166]

Bruine de Bruin, Wändi. 2006. "Save the last dance II: unwanted serial position effects in figure skating judgments." *Acta psychologica* 123 (3): 299–311. [166, 192]

Bursztyn, Leonardo, and Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *Quarterly Journal of Economics* 130 (3): 1329–1367. [162]

Campbell, Bryan, and John W. Galbraith. 1996. "Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments." *The Statistician* 45 (4): 521–526. [163]

Colombo, Luca, and Gianluca Femminis. 2008. "The social value of public information with costly information acquisition." *Economics Letters* 100 (2): 196–199. [172]

Darby, Michael R., and Edi Karni. 1973. "Free Competition and the Optimal Amount of Fraud." *Journal of Law and Economics* 16 (1): 67–88. [163]

Dohmen, Thomas J. 2008a. "Do professionals choke under pressure?" *Journal of Economic Behavior & Organization* 65 (3-4): 636–653. [166]

Dohmen, Thomas J. 2008b. "The Influence of Social Forces: Evidence From The Behavior of Football Referees." *Economic Inquiry* 46 (3): 411–424. [176]

Dulleck, Uwe, and Rudolf Kerschbamer. 2006. "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods." *Journal of Economic Literature* 44 (1): 5–42. [163]

Edward J. Green, and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52 (1): 87–100. [177]

Falk, Armin, and Florian Zimmermann. 2017. "Consistency as a Signal of Skills." *Management Science* 63 (7): 2197–2210. [165, 176]

Fehrler, Sebastian, and Niall Hughes. 2018. "How Transparency Kills Information Aggregation: Theory and Experiment." *American Economic Journal: Microeconomics* 10 (1): 181–209. [165]

Fernando, A. Nilesh, and Siddharth Eapen George. 2021. "Debiasing Discriminators: Evidence from the Introduction of Neutral Umpires." *Working paper*, [166]

Garicano, Luis, Ignacio Palacios-Huerta, and Canice Prendergast. 2005. "Favoritism Under Social Pressure." *Review of Economics and Statistics* 87 (2): 208–216. [163, 166]

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102 (1): 33–48. [162]

Gersbach, Hans, and Volker Hahn. 2012. "Information acquisition and transparency in committees." *International Journal of Game Theory* 41 (2): 427–453. [163, 165]

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *Quarterly Journal of Economics* 133 (2): 801–870. [162, 163, 165]

International Skating Union. 2010. "Communication No. 1629." [185]

International Skating Union. 2018. "Special Regulations & Technical Rules Single & Pair Skating and Ice Dance 2018." [167]

Kahneman, Daniel, Olivier Sobony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgement*. New York: Little, Brown Spark. [172]

Kim, Jerry W., and Brayden G. King. 2014. "Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring." *Management Science* 60 (11): 2619–2644. [171]

Krause, Annabelle, Ulf Rinne, and Klaus F. Zimmermann. 2012. "Anonymous job applications of fresh Ph.D. economists." *Economics Letters* 117 (2): 441–444. [166]

Lee, Jungmin. 2008. "Outlier Aversion in Subjective Evaluation: Evidence From World Figure Skating Championships." *Journal of Sports Economics* 9 (2): 141–159. [163, 166]

Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review* 97 (1): 150–168. [165]

Li, Hao, Sherwin Rosen, and Wing Suen. 2001. "Conflicts and Common Interests in Committees." *American Economic Review* 91 (5): 1478–1497. [177]

Lichter, Andreas, Nico Pestel, and Eric Sommer. 2017. "Productivity effects of air pollution: Evidence from professional soccer." *Labour Economics* 48: 54–66. [166]

Litman, Cheryl, and Thomas Stratmann. 2018. "Judging on thin ice: the effects of group membership on evaluation." *Oxford Economic Papers* 70 (3): 763–783. [163, 171]

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. "How social influence can undermine the wisdom of crowd effect." *Proceedings of the National Academy of Sciences of the United States of America* 108 (22): 9020–9025. [162]

Maida, Agata, and Andrea Weber. 2019. "Female leadership and gender gap within firms: Evidence from an italian board reform." *ILR Review*, 0019793920961995. [166]

Mas, Alexandre. 2017. "Does Transparency Lead to Pay Compression?" *Journal of Political Economy* 125 (5): 1683–1721. [166]

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99 (1): 112–145. [162]

Mattozzi, Andrea, and Marco Y. Nakaguma. 2019. "Public versus Secret Voting in Committees." *Working paper*, [163, 165]

Meade, Ellen E., and David Stasavage. 2008. "Publicity of Debate and the Incentive to Dissent: Evidence from the US Federal Reserve." *Economic Journal* 118 (528): 695–717. [162, 165]

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159 (3810): 56–63. [171]

Morris, Stephen, and Hyun Song Shin. 2002. "Social Value of Public Information." *American Economic Review* 92 (5): 1521–1534. [163, 170, 173, 197]

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101 (4): 1410–1435. [163, 166]

Pope, Devin G., Joseph Price, and Justin Wolfers. 2018. "Awareness reduces racial bias." *Management Science* 64 (11): 4988–4995. [166]

Pope, Devin G., and Maurice E. Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review* 101 (1): 129–157. [166]

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95 (3): 862–877. [163]

Prendergast, Canice. 1993. "A Theory of "Yes Men"." *American Economic Review* 83 (4): 757–770. [163]

Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination Among NBA Referees." *Quarterly Journal of Economics* 125 (4): 1859–1887. [163, 166]

Rausser, Gordon C., Leo K. Simon, and Jinhua Zhao. 2015. "Rational exaggeration and counter-exaggeration in information aggregation games." *Economic Theory* 59 (1): 109–146. [177]

Rosar, Frank. 2015. "Continuous decisions by a committee: Median versus average mechanisms." *Journal of Economic Theory* 159: 15–65. [165]

Sandberg, Anna. 2018. "Competing Identities: A Field Study of In-group Bias Among Professional Evaluators." *Economic Journal* 128 (613): 2131–2159. [163, 166, 176, 177]

Stasavage, David. 2007. "Polarization and Publicity: Rethinking the Benefits of Deliberative Democracy." *Journal of Politics* 69 (1): 59–72. [176]

Suurmond, Guido, Otto H. Swank, and Bauke Visser. 2004. "On the bad reputation of reputational concerns." *Journal of Public Economics* 88 (12): 2817–2838. [163]

Swank, Job, Otto H. Swank, and Bauke Visser. 2008. "How Committees of Experts Interact with the outside World: Some Theory, and Evidence from the Fomc." *Journal of the European Economic Association* 6 (2-3): 478–486. [165]

Swank, Otto H., and Bauke Visser. 2021. "Committees as Active Audiences: Reputation Concerns and Information Acquisition." *Working paper*, [163, 165]

Visser, Bauke, and Otto H. Swank. 2007. "On Committees of Experts." *Quarterly Journal of Economics* 112 (1): 337–372. [165]

Zitzewitz, Eric. 2006. "Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making." *Journal of Economics & Management Strategy* 15 (1): 67–99. [163, 166, 171, 176, 177]

Zitzewitz, Eric. 2014. "Does Transparency Reduce Favoritism and Corruption? Evidence From the Reform of Figure Skating Judging." *Journal of Sports Economics* 15 (1): 3–30. [166]