

Decisions, Behavior and Societal Challenges

Inaugural-Dissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften
durch die

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

Vorgelegt von

Hua-Jing Han

aus Hangzhou, Zhejiang

Bonn

2021

Dekan: Prof. Dr. Jürgen von Hagen
Erstreferent: Prof. Dr. Lorenz Götte
Zweitreferent: Prof. Dr. Matthias Wibrat

Tag der mündlichen Prüfung: 21.01.2021

For my parents
who gave me roots and wings.

Acknowledgments

I would like to thank Lorenz Götte for being the perfect first supervisor for me — and for letting me use my own strengths, supporting my ideas and plans even when my head definitely preferred the more unconventional plans and ideas to the conventional ones, and for teaching me in his apparently endless patience that — if you look close enough — you can find beauty and potential in accidents.

Thanks to Matthias Wibral who was already providing me invaluable guidance and inspiration way before he officially became my second supervisor. I not only profited greatly from his academic insights, he was also always there as a very human voice of support who would remind me of what really matters.

In my second semester, I somewhat randomly decided to take a lecture definitely not on the second semester plan: Experimental Economics, taught by Sebastian Kube. It was there where I immediately — and apparently irrevocably — fell in love with Behavioral Economics. In a musical metaphor, his lecture was likely the exposition, and from there on, my sparked passion for Behavioral Economics became the theme, or maybe even the leitmotif, of my academic years. In these years, the theme was altered, developed and combined in various ways but would always be present. This dissertation might be the recapitulation with a coda. I would like to thank Sebastian Kube for sparking my interest in Behavioral Economics, and for being there during the whole academic journey — it is an honor that he is also there at the end as a part my dissertation committee.

A huge thank you to Benson Tsz Kin Leung and Zhihao Lim for being awesome people to write papers with. Benson was also great compass to the place where microeconomic theory meets behavioral economics. Thanks to all the other kind souls who provided me academic guidance and/or fruitful discussions in my PhD years — especially Marc Dordal Carreras for being such an amazing person during my time in Berkeley.

Thanks to my friends for being the wonderful people they are. And since this is my dissertation, I want to say special thanks to Ximeng Fang, who as a friend was the Samwise Gamgee of my PhD years.

And I would like to thank my family: My parents, who made all of this possible by coming a long way and building something amazing out of nothing, for their incredible courage and their sacrifices, their unshakable belief in us, and their unconditional love. My little sister, who is the biggest gift I ever got. And our family dog Gilda who is also a true inspiration for Behavioral Economics: I have never met a being that is more “Homo Oeconomicus” than this dog. Thanks to Nitza and Karl-Heinz Speidel for being extended family since I can think.

I am incredibly lucky and privileged to have met the countless wonderful people who supported me on my way and I am infinitely grateful. A huge heartfelt thank you to all of you — you know who you are and you truly make the world a better place.

Finally, I would like to thank K., for being the most beautiful and loving challenge and for his special talent to make me smile no matter the circumstances. As a famous commercial campaign once put so wonderfully: “Here’s to the crazy ones...” and the rest is well known.

Contents

Contents	2
List of Tables	4
List of Figures	6
Publications	7
Introduction	8
1 Organ Donation and Reciprocity	14
1.1 Introduction	14
1.2 Methods	15
1.2.1 Data	15
1.2.2 Analysis	17
1.3 Results	19
1.3.1 Descriptive Statistics	19
1.3.2 Reciprocity and Organ Donation	20
1.4 Discussion and Conclusion	23
Appendices	
1.A Appendix A: Supplementary tables	25
2 Information Overload and Confirmation Bias	31
2.1 Introduction	31
2.2 Model and Intuition	33
2.3 Experimental Design	36
2.3.1 States and Information of The Guessing Task	36
2.3.2 The procedure of the guessing task	36
2.3.3 Procedural Details	41
2.4 Analysis	41
2.4.1 Data	41
2.4.2 Variables of Interest	43
2.4.3 Empirical Strategy and Hypothesis	44
2.5 Results	47
2.5.1 Preliminaries	47
2.5.2 Switching Behavior	49
2.5.3 Quantifying Bias	51
2.6 Conclusion	53

Appendices	
2.A Appendix A: The Distributions in the Guessing Task	54
2.B Appendix B: Instructions	57
3 The Dynamics of Goal Setting: Evidence from a Field Experiment on Resource Conservation	72
3.1 Introduction	72
3.2 Experimental Design	76
3.2.1 Background	76
3.2.2 HYDRAO smart shower head	77
3.2.3 Treatment assignment	78
3.2.4 Behavioral predictions	81
3.3 Analysis and Interpretation	82
3.3.1 Randomization checks	83
3.3.2 Descriptive evidence	84
3.3.3 Estimation strategy	87
3.3.4 Underlying heterogeneity	92
3.4 Robustness Checks	94
3.4.1 Sample Selection	94
3.4.2 Fraction of time spent under flashing red light	95
3.4.3 Efficacy of stand-alone feedback lights	96
3.5 Conclusion	97
Appendices	
3.A Appendix A: Supplementary figures	99
3.A.1 Floor plan of the residential colleges	99
3.A.2 Posters	100
3.B Appendix B: Supplementary tables	102
4 Bibliography	103

List of Tables

1.1	Descriptive statistics	17
1.2	Possession of a donor card and reciprocity	21
1.3	General willingness to become an organ donor and reciprocity.	22
1.4	Variable definitions	25
1.5	Correlations of personality variables, sample included in the analyses	26
1.6	Possession of a donor card and reciprocity, using standardized factors from reciprocity items	27
1.7	General willingness to become an organ donor and reciprocity, using standardized factors from reciprocity items	28
1.8	Possession of a donor card and reciprocity, using combined samples from 2006 and 2009 and reduced reciprocity measures	29
1.9	General willingness to become an organ donor and reciprocity, using combined samples from 2006 and 2009 and reduced reciprocity measures	30
2.1	An example of the underlying distributions and numbers for subjects in two pairs which belong to the same super-pair.	37
2.2	Frequencies and proportion of observations where 4, 5, 6 and 7 num- bers have been seen in phase 2.	42
2.3	Analysis of the (absence) of treatment effects after Phase 1, OLS.	48
2.4	Proportion of observations in which subjects have made a switch- ing mistake. Only complete pairs with the same Bayesian switching choice are included.	50
2.5	OLS of Switching Decisions after Phase 2.	50
2.6	OLS on quantified bias when numbers seen in phase 2 are in aggregate belief-challenging.	52
2.7	OLS on quantified bias when numbers seen in phase 2 are in aggregate belief-confirming.	52
2.8	Simulated proportion of observations where subjects should switch from believing “High” after phase 1 to believing “Low” after phase 2, or from believing “Low” to believing “High”, with 10,000 simulations of 2,000 observations	55
2.9	Distribution of Bayesian beliefs given 5 and 12 draws, with 10,000 simulations of 2,000 observations.	56
2.10	Distribution of Bayesian beliefs given 5 and 12 draws, with 10,000 simulations of 2,000 observations.	56
3.1	Summary of treatment assignments	79
3.2	Randomization checks	84
3.3	Effects of moral suasion and real-time feedback on water use per shower	90

3.4	Effects of moral suasion and real-time feedback on number of showers per day	91
3.5	Interaction effects with baseline water use	93
3.6	Effects of changing goals on time spent under flashing red light	96
3.7	Effect of moral suasion and real-time feedback on water use per shower using full sample of recorded showers	102

List of Figures

1.1	Distribution of positive and negative reciprocity	19
2.1	The two distributions shown in bar charts.	36
2.2	Sequence of a round.	38
2.3	The belief elicitation screen.	39
2.4	A screen shot of Phase 1.	39
2.5	A screen shot of phase 2 in the treatment condition.	40
2.6	A screen shot of phase 2 in the control condition.	40
2.7	Distribution of the absolute difference between elicited and Bayesian belief.	48
2.8	Scatter plot and regression line with Bayesian belief on x-axis and Elicited belief on y-axis.	49
2.9	Logarithmic Odds ratios of the numbers 1-8	55
3.1	Experimental site	77
3.2	Experimental groups in phase 1 vs phase 2.	79
3.3	Implementation of <i>18L GOAL</i> group	80
3.4	Distribution of baseline water use per shower	83
3.5	Extensive margin of shower behavior by experimental groups	85
3.6	Intensive margin of shower behavior by experimental groups	86
3.7	Average treatment effects by experimental groups	87
3.8	Typical floor plan of Cinnamon and Tembusu colleges	99
3.9	Posters for the <i>Moral Suasion</i> group in each phase	100
3.10	Posters for the <i>18L GOAL</i> group in each phase	100
3.11	Posters for the <i>28L GOAL</i> group in each phase	101

Publications

Parts of this dissertation have been published as articles. The published articles can be found online using the following references.

Chapter 1

Han, H.J., and Wibrals, M. (2020) Organ donation and reciprocity. *Journal of Economic Psychology*, 81, 102331.
<https://doi.org/10.1016/j.joep.2020.102331>

Chapter 2

Goette, L., Han, H.J., & Leung, B.T.K. (2020) Information Overload and Confirmation Bias. *Working paper*.
<https://doi.org/10.17863/CAM.52487>

Chapter 3

Goette, L., Han, H.J., & Lim, Z.H. The Dynamics of Goal Setting: Evidence from a Field Experiment on Resource Conservation. *In preparation*.

Introduction

We have set foot on the moon, explored deep seas and the highest mountains, we can solve complex equations and program elaborate algorithms, we have even decoded the human genome; but when it comes to our decisions, or our decision making, or our behavior, we humans are still a puzzle sometimes. And we live in turbulent, complicated times: Polarization is driving societies apart, we need to solve the challenges of climate change and resource scarcity to keep our planet habitable for the next generations,... but with all these pressing issues with no easy answers, there is also hope. Hope that we can manage to solve these challenges and make things better.

This thesis has two aims: First, to contribute to the scientific understanding of human behavior and decisions. One reason why human behavior can be so puzzling is that human judgments and decisions are influenced by innumerable factors. Some of these factors are our preferences, our beliefs, and our aspirations and goals. This thesis addresses each of these factors in separate chapters. Second, to apply the insights into human behavior and decisions to solve real-world problems, and to show the potential of behavioral research to tackle societal challenges such as the shortage of donor organs; confirmation bias which can lead to polarization; and resource scarcity.

To do so, this thesis draws from the full range of Behavioral Economic tools and methods, from survey data over laboratory experiments to field experiments.

The first topic is the role of preferences, especially reciprocity, on organ donation attitude and behavior. In 2019, there was a big debate in German politics about whether to introduce the opt-out solution. The final decision of the German parliament was against the opt-out solution, which would by default assume that everyone has agreed to become an organ donor post-mortem until indicated otherwise. Instead, it adopted a law keeping the opt-in system, but putting a stronger emphasis on engaging citizens with the topic of organ donation and the possible actions. Put more broadly, the question about willingness to become an organ donor and the possession of a donor card is about public good provisions in the health sector — and quite literally a matter of life and death.

Chapter 1 uses a representative data set from the German Socioeconomic Panel (SOEP) to examine the association between reciprocity, i.e. the inclination to punish unkind acts or to reward kind ones, and organ donation willingness, as expressed in the general willingness to become an organ donor, as well as behavior, as expressed in having signed a donor card. If the willingness to donate organs is perceived as kind, people with a higher positive reciprocity could be more likely to reciprocate the (prosocial and kind) willingness of others with a higher willingness themselves (given equal beliefs). In the same line of thought, if the unwillingness to donate organs

is perceived as unkind, people with a higher negative reciprocity might “punish” the prevalence of antisocial attitudes with a lower willingness to become an organ donor (given equal beliefs). The same reasoning might apply to organ donation behavior, i.e. the possession of a donor card. We find that there is a significant negative correlation between negative reciprocity and organ donation behavior. For organ donation attitude, we find a significant positive correlation between positive reciprocity and organ donation attitude, as well as a significant negative correlation between negative reciprocity and organ donation attitude. These results have interesting policy implications as they show that there are promising policy channels beyond the popular and well known defaults which can be leveraged to increase organ donation willingness on the one hand, and to get people to act on their willingness on the other hand. One possibility would be to increase the visibility and salience of positive examples. Another possibility might be to help people not to underestimate the prosocial attitude and behavior of the general population (since an underestimation might trigger negative reciprocity while a high belief would trigger positive reciprocity instead). Our results show that policies which tackle the shortage of donor organs — and other public goods — from a holistic societal perspective might be promising.

The second topic is belief formation, information overload and confirmation bias, with a focus on how information overload affects confirmation bias, the tendency of humans to seek or interpret evidence in ways that affirm one’s existing beliefs, expectations, or a hypothesis at hand (Nickerson, 1998). With the advent of the internet came a major change of how information can be accessed: We have all the information from the entire world wide web at our fingertips. However, this means that the amount of information has increased drastically — but our cognitive capacities have not. So, the internet in particular can be a curse and a blessing: Information is available for everyone and thus, much more democratic than it was, but at the same time, misinformation is a much greater problem and some mechanisms on the internet can drive cognitive biases in information processing. Especially the amount of information available can be overwhelming, as can the frequency in which we are confronted with ever new pieces of information. Imagine scrolling through a news feed. As it is cognitively impossible to process all information, people have to choose which information they process (for example by clicking on the article and reading it). Confirmation bias is a phenomenon where people under-react to belief-challenging pieces of information (for example by ignoring them) and over-react to belief-confirming information. This is especially dangerous when there is misinformation which confirm existing beliefs since it can be tempting to process them and ignore the truthful facts instead. Especially on the internet, misinformation is quite ubiquitous (see e.g. Allcott and Gentzkow, 2017; Kata, 2010). Combined with confirmation bias, this can lead to echo chambers — a phenomenon on social media which can delete the common ground for any fruitful fact-based discussions (see e.g. Del Vicario et al., 2016) and further give rise to ingroup-outgroup dynamics, adding fuel to the fire. All of these phenomena are examples for what can drive polarization, which is a major challenge for modern societies and democracies (see e.g. McCoy et al., 2018). The World Economic Forum has named “domestic political polarization” as one of the top risks in 2020 (World Economic Forum, 2020), and it is more important than ever to understand the behavioral mechanisms behind it.

Chapter 2 examines the link between information overload and confirmation bias in a laboratory experiment. We define confirmation bias as an asymmetric belief updating behavior. A stronger information overload might lead to a stronger confirmation bias if subjects update their beliefs more with belief-confirming information and if they under-react to belief-challenging information (by updating less) in an environment where information overload is more severe (aka the amount of information is cognitively more taxing).

In our experimental setting, subjects receive a sequence of numbers, which are drawn from either a “low” or “high” distribution. We call these distributions “computers” to make it more intuitive for the subjects. While the “low” distribution is more likely to generate small numbers, the “high” distribution is more likely to generate large numbers. Subjects have to navigate through the sequence within a time limit of 30 seconds and then report their beliefs on which computer generated the numbers they have seen. In each round of the experiment, two subjects are matched and assigned to either the treatment or control condition. We impose a stronger information overload in the treatment than in the control condition: We hold constant the available signals for the matched subjects in the treatment and control conditions but vary the difficulty of belief-updating to isolate the effect of information overload on belief updating.¹ In the treatment condition, subjects navigate through the sequence by clicking the “next” button and they only see one signal at a time; while in the control condition, they advance through the sequence only when their matched subject in the treatment condition has clicked the “next” button. The important distinction is that as the control subjects advance through the sequence, the preceding numbers remain visible and they observe multiple numbers at the same time. For the treatment subjects, the preceding number disappears when they display the next number — this induces a higher information overload since it is cognitively more taxing.

We find that subjects in the treatment condition under-react to belief-challenging information while their reaction towards belief-confirming information is unaffected by information overload. This asymmetric bias holds for their switching attitude, i.e. how likely they are to make mistakes in their decisions to switch their belief between the “high” and “low” distribution, as well as for the quantified bias, i.e. the discrepancy between the subjective belief updating of the subjects and the Bayesian benchmark.

These findings have important policy implications: First, they illustrate that information overload as for example on the internet can indeed give rise to confirmation bias — and promote polarization even though information is more easily accessible. Further, our findings show that more information is not always better — on the contrary, more information might lead to a stronger confirmation bias and thus, even stronger polarization. This implies that smart policies to combat polarization on the internet have to consider these effects, explore other channels and be carefully designed. Just providing even more information in the futile hope that people might consume the information, process it and update their beliefs accordingly can lead to unintended side-effects. For successful policies to battle polarization and misinformation, it is important to take into account our human cognitive constraints.

¹Since we hold the available signals constant in the pair of matched subjects, a Bayesian subject will form the same belief in both conditions.

The third topic is the dynamics of goal setting and how goals and real-time feedback can help resource conservation. Shower consumes water and energy, but both are limited resources where excessive consumption might cause serious problems for environment and society. We will focus on water here: Water — as ubiquitous as it may seem — might become scarce. Factors like climate change, population growth and economic development have placed an increasing stress on the global water supply. The Global Risk Report 2020 of the World Economic Forum lists “water crises” as one of the top 10 risks in terms of likelihood and impact (World Economic Forum, 2020). Burek et al. (2016) estimate that 3.6 billion people (comprising 51% of the global population) worldwide are living in areas which face the potential of severe water scarcity, and this figure is set to increase to between 4.8 and 5.7 billion by 2050. In particular, the slate of water-related challenges will likely be most acute in Asia, currently home to 73% of the affected people.

Goal setting might play an important role in resource conservation as in many domains, marginal costs are low for individuals, as is often the case for water or electricity with certain rental agreements, and monetary incentives might not be feasible. We study the role of goal setting and real-time feedback in a setting where marginal costs for individuals are zero. From a behavioral perspective, goal setting alone is already an important topic since goals are a ubiquitous tool for performance management and motivation, in the private as in the public sector. However, setting goals can have unexpected side-effects. In economic literature, goals are often viewed as inheriting the properties of reference points (Heath et al., 1999; Kahneman and Tversky, 1979; Tversky and Kahneman, 1991), which can cause individuals to experience loss aversion and diminishing sensitivity around goals — further, their old performances might also influence the setting of new goals. As goals often change over time, for example in the form of business objectives which adapt due to changes in the economic environment, it is important to understand the dynamics of goal setting, goal difficulty and the effects when goals are changed.

Chapter 3 examines the dynamics of goal setting in a randomized field experiment, applied to resource conservation in the shower, to analyze how residents respond to different (exogenous) goals on shower water use over time.

The field experiment is set in two residential colleges at the National University of Singapore with more than 600 students in total. We use moral suasion in the form of posters in the shower cubicles to set goals under which the water usage should be kept and HYDRAO smart shower heads to provide real-time feedback. These smart shower heads measure water consumption during the shower and feature a LED light which changes the color based on water usage in real-time.

Our experiment had three stages: First, a baseline period to collect information on the pre-experimental shower behavior of the subjects, followed by phase 1 and 2 with four experimental groups each: a control group, and three treatment groups (“Moral Suasion”, “18L”, “28L”). The assignments to the treatments groups were permanent for the experiment. The control group received no shower poster and also no real-time feedback in any phase of the experiment. The “Moral Suasion” group received a shower poster with a specific level under which the water use should be kept. This water level (the goal) was 28L for phase 1, and 24L for phase 2. The other two treatment groups were to show the dynamics of goal setting: In phase 1, the 18L group started with a hard goal, while the 28L group started with a moderate goal. For phase 2, the goals for both groups were changed to the same intermediate

goal of 24L.

In detail, the 18L group received a shower poster with an 18L goal for phase 1 and corresponding real-time feedback via the smart shower heads, while the 28L group received a shower poster with a 28L goal for phase 1 and corresponding real-time feedback via the smart shower heads. For phase 2, the 18L and 28L group were moved to a 24L goal with the corresponding posters and real-time feedback.

This setting allows us to test central predictions derived from the (behavioral) economics literature:

First, we can test whether an ambitious goal in our setting (the 18L goal) elicits strong effort because the individual is so far in the loss-domain that reducing the loss has a high marginal value, or whether the ambitious goal lowers the effort because of diminishing sensitivity. Second, we can test what happens when both, the hard 18L goal and the moderate 28L goal, are moved to the intermediate 24L goal: One prediction could be that this could lead to the same reference point and thus, the same outcome. However, if reference points are shaped by “new” expectations or lagged outcomes, we would observe different effects depending on the former goal: Then, changing from the 18L to the 24L goal would decrease effort because the now easier 24L goal moves subjects into the gain domain, reducing the marginal value of saving water, while changing from the 28L goal to the 24L goal would increase efforts. Third, by converging the 28L goal to a 24L goal, we can test whether reference dependence makes a gradual increase of goal difficulty more effective.

We find that moral suasion alone does not have significant effects on water savings. But paired with real-time feedback, the effects on conservation efforts are large and significant. Interestingly, in phase 1, we do not find differences in performance between the 18L and the 28L group. But in phase 2, when the goals for both groups are changed to the same 24L goal, the 18L group then performs worse, while the 28L group increases their performance relative to phase 1.

When looking deeper and including interactions with the baseline use, we find that the seemingly same performance in phase 1 between 18L and 28L groups masks an underlying heterogeneity: While there is only a small interaction effect between baseline use and reaction for the 18L group in phase 1, the effect for the 28L group is highly significant. Fascinatingly, this heterogeneity in interaction with baseline use carries over to phase 2, when both groups are moved to the same goal.

These findings point to the possibility that initial goal difficulty might not show immediate effects but create lasting effects which show only when goals are changed. Furthermore, there might be permanent effects of initial goal setting which goes beyond our predictions as we see in the heterogeneity in interaction with baseline use. Our findings show that it is highly important to select goals and the optimal level of goal difficulty carefully since a “suboptimal” level of difficulty might backfire even when it is adjusted later. They are also highly relevant for policies as they show that behavioral interventions can be valuable tools to tackle the societal challenge of resource scarcity even without any monetary incentives.

Taken together, this thesis examines three factors influencing human behavior, judgments and decisions, and applies these insights to tackle societal challenges.

For policy and decision makers, this thesis can be a warning and encouragement alike:

The warning is that policies and measures can backfire when not wisely chosen

and designed. There simply is no magic bullet, no panacea. For example, goals are a popular and widely used tool for performance management, but they can have serious side-effects: Too ambitious goals can apparently lead to a suboptimal performance and create permanent effects which also stay when goals are adjusted later (Chapter 3). And in the battle against misinformation, the provision of true information and facts seems like a logical measure. However, these measures have to be designed carefully since simply providing more information can lead to information overload and thus give rise to (even stronger) confirmation bias. Thus, one has to consider how the information is provided and take human cognitive constraints into account (Chapter 2).

But most importantly, this thesis should provide hope and encouragement: In our time, we are facing numerous societal challenges which we urgently need to solve. We do have tools to do so and we have by far not realized the full potential of behavioral tools in policy making. Holistic policy measures to increase organ donation, which take people's social preferences such as reciprocity and their beliefs into account, might yield double dividends; promising channels might be to increase visibility and salience of positive examples (Chapter 1). A better understanding of channels which might give rise to confirmation bias can help us to design policies to prevent this — and help tackling the societal challenge of polarization (Chapter 2). And goal-setting and real-time feedback are powerful tools to increase efforts in resource conservation which is a vital part of combating climate change. As a behavioral intervention in our setting, they even worked without any monetary incentives and in a setting where the marginal costs of resource consumption were zero (Chapter 3). All three chapters show that an in-depth understanding of the drivers and influencing factors behind human behaviors and decisions is already valuable. Moreover, there are easily applicable behavioral tools and interventions that we can derive from these insights into human behavior which can be used to tackle the societal challenges of our time.

Chapter 1

Organ Donation and Reciprocity

JOINT WITH MATTHIAS WIBRAL

1.1 Introduction

In June 2019, there were 113,325 people on the waiting list for an organ transplant in the US. On average, 17 people from the waiting list died every day in 2018 (<https://optn.transplant.hrsa.gov/>). In the effort to increase the supply of suitable donor organs a better understanding of the determinants of organ donation attitudes and behavior is crucial.¹ Personality has been an important focus of research in this endeavor (e.g. Bekkers, 2006; Demir and Kumkale, 2013; Hill, 2016). In this paper, we contribute to this strand of research by studying a potential relation between negative or positive reciprocity, that is, the inclination to punish unkind acts or to reward kind ones, and organ donation attitudes and behavior.

A relation between positive reciprocity and organ donation behavior seems intuitive if people regard others' willingness to donate organs as kind. For equal beliefs about other people's organ donation attitudes and behavior (see Methods for a discussion), people with higher positive reciprocity should then be more willing to donate organs. Anecdotal evidence for this comes, for example, from the so-called "Nicholas Green effect", named after an American boy who was killed by robbers while vacationing in Italy. His parents consented to donating his organs to patients awaiting transplantation. This choice received intense media coverage, and contributed to a threefold increase in organ donations nationally (Redelmeier and Woodfine, 2013). A study by the Behavioral Insights Team (2013) also provides indirect evidence that positive reciprocity might play a role in organ donation behavior. The study investigates the effect of different messages on a British government website on subsequent registration as an organ donor. An appeal to positive reciprocity ("If you needed an organ transplant, would you have one? If so please help others.") worked better than other messages and lead to substantial increases in registrations. In a different health related context, Rönnerstrand and Sundell (2015) found that people were more willing to postpone antibiotic treatment for the sake of limiting overuse when the doctor stated that other individuals were also willing to do so.

Analogously, unwillingness to donate organs could be viewed as unkind, especially if coupled with a willingness to accept donor organs. Given equal beliefs,

¹We refer to the general willingness to donate organs post-mortem as organ donation attitude, and to a specific action taken in this regard (i.e., signing a donor card), as organ donation behavior.

people with higher negative reciprocity should then be less willing to donate organs. Negative reciprocity towards free riders, that is, people who are willing to accept an organ but refuse to donate one, was considered an important motive for low donation rates in Israel (Lavee et al., 2010). Israel has since introduced a new policy which gives priority on organ donor waiting lists to those who have previously registered as organ donors. Donation rates rose considerably after the introduction of this policy (Stoler et al., 2017).² However, it is not clear to which degree this is due to the introduction of these incentives, the lower scope for negative reciprocity, or other factors. Siegel et al. (2016) provide further indirect evidence for the importance of negative reciprocity for organ donations. They found that negative experiences and negative emotions during the visit to the Department of Motor Vehicles where the registration as an organ donor typically occurs in the US lead to lower willingness to register as an organ donor.

While previous findings thus suggest that both positive and negative reciprocity could play an important role for organ donations, direct evidence on such a relation is missing. Establishing a correlation between reciprocity, and organ donation attitudes and behavior is an important first step towards an understanding of why policy measures targeting positive or negative reciprocity could be successful. In this paper, we therefore use a unique data set from the German Socioeconomic Panel (SOEP) to study this correlation in a representative sample of the German population.

1.2 Methods

1.2.1 Data

We use data from the SOEP Pretest in 2009 (see <https://www.diw.de/en/soep> for more information on the SOEP). The SOEP pretest is representative for the population in Germany aged 16 and older. This data set is unique because it contains measures of organ donation attitudes and behavior as well as reciprocity. We describe each of these in turn.

At the time of the survey, Germany had the extended consent solution: a deceased person must carry a donor card consenting to organ donation to become an organ donor post-mortem, otherwise, their next of kin will decide. Families have been shown to act in accordance with the preferences of the deceased when this information is known (Martinez et al., 2001; Radecki and Jaccard, 1997). As a considerable number of those who do not have an organ donor card nevertheless mention their preference to their family, a post-mortem donation from a person who was willing to donate but had no donor card is more likely than from someone who was unwilling to donate. It is therefore also interesting to study general willingness to donate and not just possession of a donor card.

In the SOEP Pretest, participants were first asked if they were generally willing to be an organ donor post-mortem (yes/no question; “general willingness”). Then, those who answered with “yes” were asked if they had a donor card (yes/no question; “card possession”).³ In Germany, it is also possible to have a donor card and

²Two laboratory experiments, Herr and Normann (2016) and Herr and Normann (2019), also found that giving priority to registered donors lead to higher registration rates.

³The exact wordings of the survey questions were: “*Are you basically willing to donate organs*”

explicitly say on this card that one is *not* willing to be an organ donor. In our sample, this group is included in the “not willing” group.

The reciprocity measures in our data set (Richter et al., 2013) are based on the scale developed by Perugini et al. (2003). They have been widely used and linked to important real world labor market behavior and life outcomes (see, e.g., Caliendo et al., 2012; Dohmen et al., 2009; Fehr, 2009; Knoch et al., 2006). Positive reciprocity is defined as the inclination to respond to kind actions with kind behavior. Negative reciprocity is defined as the inclination to respond to unkind actions with punishment (Fehr and Gächter, 2000). It is important to point out that, although potentially related, positive and negative reciprocity are different concepts, and not merely two sides of the same coin (Dohmen et al., 2009; Yamagishi et al., 2012). This can also be seen in table 1.5 which shows the pairwise correlations of our personality variables.

Each reciprocity measure consists of 3 items which could be answered on a 7-point Likert scale. The items for negative reciprocity are:

- If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost.
- If somebody puts me in a difficult position, I will do the same to him/her.
- If somebody offends me, I will offend him/her back.

The items for positive reciprocity are:

- If someone does me a favor, I am prepared to return it.
- I go out of my way to help somebody who has been kind to me in the past.
- I am ready to assume personal costs to help somebody who helped me in the past.

Following Dohmen et al. (2009), we construct our measures of positive and negative reciprocity by averaging the responses over the 3 items (positive reciprocity: $\alpha = 0.64$; negative reciprocity: $\alpha = 0.80$).⁴

An additional advantage of the data set is that it enables us to control for a wide range of demographic and personality variables that have been related to organ donation attitude and behavior in previous work. The controls included are: age, level of education, marital status, sex, and household income per capita, as well as attitude towards risk (Dohmen et al., 2011), and the Big 5 personality traits. The latter were assessed using a 16-item Big Five inventory (Richter et al., 2013) that was developed for the German SOEP. Table 1.4 in the appendix presents definitions for all variables. Table 1.1 shows the descriptive statistics for all variables for the entire SOEP Pretest 2009 sample and the final sample included in our analyses (i.e., the subjects which answered all questions on the variables we included).⁵

after death?” and “Have you determined your willingness to donate organs in the written form of an organ donor card?”

⁴We also conducted a factor analysis on the reciprocity items as a robustness check: Two factors emerge from the analysis, with the positive reciprocity items loading high on one factor, and the negative reciprocity items loading high on the other factor. Using the factors instead of the averages for our analyses yields very similar results. We report these below.

⁵Non-parametric tests on the socio-demographic variables comparing those participants included in the analyses to those who are not indicate that the sample included in the analyses is significantly more male. For the other socio-demographic variables the differences are not significant.

Previous findings (Bekkers, 2006; Hill, 2016; van Andel et al., 2016; Wakefield et al., 2010; Walkowitz, 2019) also emphasize the importance of controlling for altruism. We do so by including a dummy variable indicating whether the respondent donated money in the last months.⁶ Finally, another unique feature of our data is information on whether the respondent has drafted an advance directive, that is, a legal document specifying what actions should be taken if she is no longer able to make decisions for herself. Our final sample consists of 683 individuals who answered all questions on organ donation, positive and negative reciprocity, and the control variables.

Table 1.1: *Descriptive statistics*

	SOEP Pretest 2009, entire sample			Included in the analyses		
	Mean	SD	Obs.	Mean	SD	Obs.
Card Possession	0.12	0.32	946	0.13	0.34	683
Willingness	0.46	0.50	946	0.49	0.50	683
Positive Reciprocity	5.96	0.90	986	5.90	0.91	683
Negative Reciprocity	2.97	1.53	976	3.06	1.55	683
Risk	4.45	2.65	1000	4.58	2.54	683
Agreeableness	5.49	1.02	993	5.43	1.00	683
Openness	4.89	1.17	975	4.88	1.14	683
Conscientiousness	5.91	0.93	984	5.83	0.95	683
Neuroticism	3.81	1.24	993	3.85	1.21	683
Extraversion	4.99	1.15	992	5.02	1.15	683
Altruism	0.49	0.50	998	0.51	0.50	683
Advance Directive	0.13	0.34	1001	0.14	0.35	683
Age	52.13	18.59	1002	51.84	18.23	683
Male	0.44	0.50	1007	0.46	0.50	683
Income per capita	1.06	0.67	780	1.08	0.69	683
Education	2.02	1.11	966	2.05	1.11	683
Ever married	0.78	0.41	999	0.77	0.42	683
Total Observations			1007			683

1.2.2 Analysis

Using logit regressions, we investigate two dimensions in which reciprocity could manifest itself, organ donation behavior and attitude. To study the former, we regress possession of a donor card on our reciprocity variables. To study the latter, we regress general willingness to become an organ donor (independent of whether this has been put into practice by signing a donor card) on our reciprocity variables. In both cases, we add personality, and socioeconomic variables as controls in a stepwise fashion.⁷

⁶Respondents were also asked how much they would donate out of an unexpected gift of 10.000€. Using this variable as an alternative control for altruism does not change the qualitative results.

⁷Our main results also hold if we include all subjects who have answered the questions which are relevant for the respective specification (N = 919 for the regressions only including positive and

Before we present our results, two aspects of the analysis merit discussion. First, the influence of positive or negative reciprocity as a trait or preference on organ donation attitudes and behavior is likely to also depend on (beliefs about) others' behavior. Our analysis therefore relies on additional identifying assumptions concerning what individuals condition their reciprocity on. Here we discuss two plausible (and not mutually exclusive) assumptions.

First, it could be the case that reciprocal behavior conditions on the average behavior in the population, or more precisely, on beliefs about the average behavior in the population, similar to findings in public good games. For example, Croson (2007) finds a significant and positive relationship between an individual's own contribution and their beliefs about the contributions of others in their group. Between two people who score equally on our measure of negative reciprocity, the one with the lower belief about average willingness to donate organs should then *ceteris paribus* be less likely to be willing to donate their organs. Under the assumption that all individuals hold the same beliefs our model correctly estimates the correlations between the two reciprocity measures, and organ donation attitude and behavior. However, data from a survey by the German Federal Centre for Health Education in 2014 show quite some heterogeneity in beliefs in the population (Caille-Brillet et al., 2015).⁸ For heterogeneous beliefs, our model will still correctly estimate the *average* correlations between reciprocity, and organ donation attitude and behavior if beliefs and reciprocity are uncorrelated. Unfortunately, SOEP pretest participants were not asked about their beliefs about others' general willingness to donate organs or their possession of a donor card. We therefore cannot directly incorporate these beliefs into our main analysis. However, to shed more light on a possible correlation between beliefs about organ donation and reciprocity, we conducted a survey among 193 students of the University of Bonn who were participants in an unrelated experiment in October 2017. In addition to the positive and negative reciprocity items, they were asked about their belief about which percentage of the adult German population possesses an organ donor card. We find that beliefs about the percentage of organ donor card holders in Germany are neither correlated with positive reciprocity (Spearman's $\rho = 0.046, p = 0.529$) nor negative reciprocity ($\rho = -0.134, p = 0.063$).⁹ Our findings should nevertheless be interpreted with the caveat in mind that the student sample might not be representative for the correlations between reciprocity and beliefs in the general population.

A second plausible identifying assumption is that people react reciprocally to particularly salient examples of behavior, for example, the case of Nicholas Green discussed in the introduction. This would imply that people who score higher on positive reciprocity would show a stronger reaction to news like the one about Nicholas

negative reciprocity, $N = 890$ for the regressions including reciprocity and personality variables).

⁸In this survey, participants were asked "What do you think how many people in Germany are currently willing to donate their organs?". The survey only gives four answer options: "more than 50%" (4% of participants chose this option), "about 50%" (10% of participants), "less than 50%" (50% of participants), and "only a few" (36% of participants). It is also quite striking that participants substantially underestimate the general willingness to donate organs which in the same sample was 71%.

⁹Similar to the belief data reported in the previous footnote, participants in our sample substantially underestimated the percentage of the population that possesses an organ donation card (average estimate: 24%; percentage reported by the Federal Centre for Health Education in early 2018: 36%).

Green than those who score lower on positive reciprocity. Evidence from Germany suggests that this also works in the opposite direction. In 2012, investigations revealed that doctors in several transplantation centers had manipulated waiting lists for donor organs. Survey data from the Federal Centre for Health Education show that general willingness dropped by four percentage points and donor card possession by three percentage points from 2010 to 2012 (Schmidt et al., 2013). In 2013, respondents were asked whether they had ever changed their opinion about organ donation. 37% of those that replied with “yes” stated that they had done so because of the scandal (Schmidt et al., 2014). Overall, 60% and 53 %, respectively, of those not willing to donate organs stated potential abuse through organ donation trade, and concerns about the fair distribution of organs as their reasons. Under the identifying assumption that people react reciprocally to particularly salient examples of behavior, our model correctly estimates the correlations between the two reciprocity measures, and organ donation attitude and behavior if all individuals are exposed to the same salient examples which seems plausible given widespread media reports.¹⁰

A final concern about our analysis might be multicollinearity. Table 1.5 in the appendix shows the pairwise correlations between our personality variables. However, tests to see if the personality variables met the assumption of collinearity indicated that multicollinearity was not a concern (all VIFs < 1.41).

1.3 Results

1.3.1 Descriptive Statistics

Out of the whole sample, 46% answered that they are generally willing to become a post-mortem organ donor. Out of these 46% who are generally willing to become an organ donor, 26% stated that they have a donor card. This means that 12% of the entire sample have a donor card documenting their willingness to become an organ donor.

Figure 1.1 shows the distribution of the reciprocity measures in our sample.

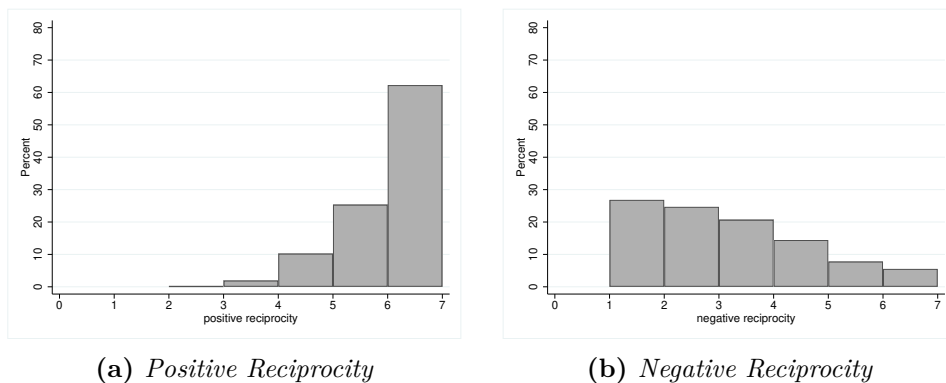


Figure 1.1: *Distribution of positive and negative reciprocity*

Positive reciprocity has a mean of 5.94 and a standard deviation of 0.91, the distribution is left-skewed. Moderate to strong positive reciprocity is the norm in

¹⁰Our model also correctly estimates the average correlations if exposure intensity is uncorrelated to reciprocity.

our sample. We observe more variation for negative reciprocity. The mean is 2.99 with a standard deviation of 1.53, and the distribution is right-skewed. The modal response is complete disagreement with all three negative reciprocity items, but a substantial number choose intermediate values or even complete agreement. Our results are in line with those found in an even bigger sample from the German population (Dohmen et al., 2009). For our regression analyses, we use standardized reciprocity measures, as well as standardized measures of the Big 5, and attitude towards risk.

1.3.2 Reciprocity and Organ Donation

Table 1.2 shows the results of our logit regression of card possession (i.e., generally willing to donate and in possession of a donor card) on our reciprocity variables. Our reciprocity measures and all personality control variables are standardized. We report the odds ratios for an increase of one standard deviation in the respective variable. We find that the coefficient for positive reciprocity is not significant (odds ratio [OR] 1.033, $p = 0.776$). However, the coefficient for negative reciprocity is significant: scoring higher on negative reciprocity is associated with a lower likelihood of card possession (OR 0.769, $p = 0.013$). These results still hold after including our wide range of control variables (positive reciprocity: OR 0.875, $p = 0.321$; negative reciprocity: OR 0.756, $p = 0.038$).

Our second analysis concerns the relation between reciprocity and general willingness to become an organ donor. The results of our logit regression of general willingness on our reciprocity variables are shown in Table 1.3. We find highly significant coefficients of both positive and negative reciprocity: While a higher score in positive reciprocity is associated with a higher likelihood of being generally willing to become an organ donor (OR 1.409, $p < 0.001$), a higher score in negative reciprocity is negatively associated with the general willingness to become an organ donor (OR 0.750, $p < 0.001$). Again, these results hold after controlling for personality and socio-demographic control variables (positive reciprocity: OR 1.322, $p = 0.001$; negative reciprocity: OR 0.731, $p < 0.001$).¹¹

As expected, we find a significant positive correlation of our measure for altruism with both, donor card possession (OR 2.537, $p = 0.001$), and general willingness to become an organ donor (OR 1.635, $p = 0.007$). This is in line with a previous finding (Morgan and Miller, 2002).

To the best of our knowledge, we are the first to look at a possible link between

¹¹We report two robustness checks in the appendix. First, using the factors emerging from the factor analysis instead of the averages for our analyses yields very similar results (see table 1.6 and 1.7 in the appendix). Second, the SOEP Pretest 2006 (with different participants) also contains information on general willingness to donate organs and possession of a donor card. However, only two of the three items for positive reciprocity and one of the three items for negative reciprocity were elicited. To reduce measurement error and ensure comparability with other studies using the standard reciprocity measures with 3 items each we do not include these data in our main analyses. However, including the 2006 data yields a similar pattern of results concerning reciprocity (see tables 1.8 and 1.9). In these regressions, we use the average of the two items for positive reciprocity, and the item for negative reciprocity available in 2006 as our measures of positive and negative reciprocity, respectively. The general willingness to donate organs is positively associated with positive reciprocity and negatively associated with negative reciprocity. For card possession, the odds ratio for negative reciprocity is similar in size, but not significant at the 5% level ($p = 0.066$ whereas it was $p = 0.038$ before).

Table 1.2: *Possession of a donor card and reciprocity*

	Possession of a donor card		
Positive reciprocity	1.033 (0.119)	0.953 (0.121)	0.875 (0.118)
Negative reciprocity	0.769* (0.081)	0.777* (0.093)	0.756* (0.102)
Risk attitude		0.991 (0.128)	0.921 (0.127)
Agreeableness		0.815 (0.108)	0.847 (0.116)
Openness		1.614** (0.248)	1.439* (0.242)
Conscientiousness		0.828 (0.089)	0.845 (0.101)
Neuroticism		1.040 (0.122)	1.057 (0.142)
Extraversion		1.206 (0.171)	1.073 (0.155)
Altruism		2.476*** (0.611)	2.537*** (0.685)
Advance directive			3.987*** (1.362)
Age			0.954*** (0.009)
Male			0.962 (0.264)
Income per capita in 1000 Euros			1.394* (0.231)
Education			1.226 (0.135)
Ever married			2.055* (0.746)
Constant	0.150*** (0.017)	0.075*** (0.016)	0.148*** (0.068)
Pseudo R^2	0.010	0.081	0.160
χ^2	6.370	45.196	71.197
p-value	0.041	0.000	0.000
Observations	683	683	683

Logit regressions, coefficients displayed as odds ratios; standard errors in parentheses.

Reciprocity and personality variables are standardized.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.3: *General willingness to become an organ donor and reciprocity.*

	General willingness to become an organ donor		
Positive reciprocity	1.409*** (0.112)	1.342*** (0.115)	1.322** (0.116)
Negative reciprocity	0.750*** (0.058)	0.763** (0.065)	0.731*** (0.065)
Risk attitude		1.183 (0.107)	1.114 (0.104)
Agreeableness		0.990 (0.092)	1.038 (0.100)
Openness		1.252* (0.121)	1.133 (0.117)
Conscientiousness		0.922 (0.085)	0.937 (0.089)
Neuroticism		1.041 (0.091)	1.071 (0.101)
Extraversion		1.043 (0.097)	1.009 (0.097)
Altruism		1.407* (0.232)	1.635** (0.300)
Advance directive			1.649* (0.415)
Age			0.973*** (0.006)
Male			1.054 (0.187)
Income per capita in 1000Euros			1.219 (0.151)
Education			1.068 (0.088)
Ever married			0.972 (0.236)
Constant	0.995 (0.079)	0.820 (0.096)	2.059* (0.718)
Pseudo R^2	0.038	0.061	0.096
χ^2	34.259	51.853	73.968
p-value	0.000	0.000	0.000
Observations	683	683	683

Logit regressions, coefficients displayed as odds ratios; standard errors in parentheses.

Reciprocity and personality variables are standardized.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

risk attitude and organ donation attitude or behavior. However, we do not find a significant association of risk attitude with possession of an organ donor card or the general willingness to become an organ donor.

Previous work regarding the impact of the Big 5 personality measures has found mixed results. Bekkers (2006) finds that extraversion is positively associated and neuroticism is negatively associated with having registered as an organ donor, while Demir and Kumkale (2013) find a positive association of conscientiousness and a negative association of neuroticism with organ donation intention. Hill (2016) finds no correlation between any of the Big 5 (after controlling for altruism), and organ donation attitude and intention to register as an organ donor. We replicate the latter study regarding card possession, but find a significant positive association between openness and having signed an organ donor card (OR 1.439, $p = 0.030$).

The odds ratios for our socio-demographic control variables largely replicate the findings of a meta-analysis concerning determinants of organ donation attitudes and behavior (Wakefield et al., 2010). Older respondents are both less likely to be generally willing to donate organs (OR 0.973, $p < 0.001$) and to possess a donor card (OR 0.954, $p < 0.001$, see also Mocan and Tekin (2007)). A higher income is also positively associated with card possession (OR 1.394, $p = 0.045$). While the coefficient for education is not significant for both, card possession and willingness (OR 1.226, $p = 0.065$ for card possession, OR 1.068, $p = 0.421$ for willingness), its direction is in line with previous findings (Nijkamp et al., 2008; Sehgal et al., 2016; Shacham et al., 2018).

Contrary to Mocan and Tekin (2007) and Sehgal et al. (2016), we do not find a significant association with gender. Having drafted an advance directive has a strong positive association with possession of an organ donor card (OR 3.987, $p < 0.001$); it is also positively associated with willingness to become an organ donor (OR 1.649, $p = 0.047$).

1.4 Discussion and Conclusion

We use a representative sample of the German population to study the relationship between reciprocity, and organ donation attitude and behavior. We find a significant negative correlation between negative reciprocity and the possession of a donor card. This is in line with the perception of unwillingness to donate organs as unkind. We also observe a significant relationship between positive and negative reciprocity, and the general willingness to become an organ donor. Positive reciprocity increases the probability of being willing to become an organ donor while negative reciprocity decreases this probability. Importantly, all of our main findings still hold after controlling for a wide range of socio-demographic and personality measures including altruism.

We did not have an a priori hypothesis about the relative strength of the correlations of positive and negative reciprocity with organ donation attitudes and behavior. Our results suggest that positive reciprocity is only correlated with general willingness while negative reciprocity is correlated with both general willingness and card possession. A potential explanation for this pattern of results lies in beliefs about the prevalence of kind or unkind attitudes or behavior. If people condition their reciprocity on the average attitude or behavior in the population, positive reciprocity is likely to exert a stronger influence when a kind attitude or behavior

is frequent (i.e., for general willingness to become an organ donor), and a weaker influence when it is less frequent (i.e., for signing a donor card). The role of beliefs in organ donation is therefore a promising area for future research. In addition, our findings are reminiscent of previous findings that agents tend to react stronger to negative than to positive stimuli (Baumeister et al., 2001).

The relation between reciprocity and organ donation behavior is also interesting from a policy perspective. If individuals condition their reciprocity on average behavior in the population, our findings open up the possibility of a double dividend of measures that increase organ donations, for example, publicity campaigns. These could yield an additional increase of organ donations via a feedback loop through reciprocity. If individuals condition their reciprocity on particular events, our findings could explain why (and especially with whom) reciprocity based interventions work as a study by Behavioral Insights Team (Behavioral Insights Team, 2013) shows. Another policy implication in this case is that salient reporting on cases like the one of Nicholas Green described in the introduction could be a very effective tool to increase organ donations.

These measures could be especially relevant for countries which have decided against an opt-out rule such as Germany did recently. Instead the German parliament adopted a new law that basically keeps the current opt-in system, but puts a much stronger emphasis on engaging citizens with the topic of organ donation and on registering their decision. For example, the authorities now have to hand out information material and donor cards when people get an ID, knowledge about organ donation will be part of the first aid courses required to get a driver's license, and doctors are encouraged to discuss the topic of organ donation with their patients.

All of these measures provide avenues for interventions that appeal to reciprocity. Given that people seem to underestimate the general willingness and percentage of card holders in the general population, including these figures in the information materials would be a very simple policy measure that could increase the effect of positive reciprocity and reduce the effect of negative reciprocity. In addition, salient examples could also be used in these materials, for example, the case of the German president Frank-Walter Steinmeier who in 2010 donated a kidney to his wife.

All this comes with the caveat that we are reporting correlations. Our results are a starting point for further research. We believe that especially manipulating beliefs about organ donations or using salient prompts while at the same time measuring reciprocity are promising directions. Finally, our finding of a strong correlation between having drafted an advance directive and possession of an organ donor card points to another potential policy lever. Encouraging people to draft an advance directive could have the side effect of increasing organ donations, and might be easier than directly encouraging organ donations as people have a self-interest in drafting an advance directive.

1.A Appendix A: Supplementary tables

Table 1.4: *Variable definitions*

Variable	Definition
Card Possession	Dummy: 1 = Has signed an organ donation card
Willingness	Dummy: 1 = Generally willing to become organ donor
Positive Reciprocity	Averaging 3 items, single items on a scale from 1 to 7
Negative Reciprocity	Averaging 3 items, single items on a scale from 1 to 7
Risk	Willingness to take risks: on a scale from 0 (low) to 10 (high)
Agreeableness	Averaging 3 items, single items on a scale from 1 to 7
Openness	Averaging 4 items, single items on a scale from 1 to 7
Conscientiousness	Averaging 3 items, single items on a scale from 1 to 7
Neuroticism	Averaging 3 items, single items on a scale from 1 to 7
Extraversion	Averaging 3 items, single items on a scale from 1 to 7
Altruism	Dummy: 1 = Donated money last year
Advance Directive	Dummy: 1 = Has drafted an advance directive
Age	Age in years, minimum age for participation in survey: 18
Male	Dummy: 1 = male
Income per capita	Household income per capita in 1000€
Education	Level of school leaving certificate (1-4)*
Ever married	Dummy: 1 = ever married

* 1 = Volksschul-/ Hauptschulabschluss (lower secondary school), 2 = mittlere Reife/Realschulabschluss (intermediate secondary school), 3 = Fachhochschule/Fachoberschule (certificate of aptitude for specialized short-course higher education), 4= Abitur/ Hochschulreife (upper secondary school degree giving access to university studies)

Table 1.5: *Correlations of personality variables, sample included in the analyses*

	Positive reciprocity	Negative reciprocity	Risk	Agreeableness	Openness	Conscientiousness	Neuroticism	Extraversion	Altruism
Positive reciprocity	1.00								
Negative reciprocity	-0.10**	1.00							
Risk	0.08*	0.05	1.00						
Agreeableness	0.24***	-0.33***	-0.10**	1.00					
Openness	0.22***	-0.14***	0.33***	0.16***	1.00				
Conscientiousness	0.25***	-0.17***	-0.05	0.34***	0.17***	1.00			
Neuroticism	-0.07	0.07	-0.13***	-0.07	-0.08*	-0.10**	1.00		
Extraversion	0.21***	0.02	0.24***	0.03	0.39***	0.20***	-0.19***	1.00	
Altruism	0.10**	-0.17***	0.05	0.05	0.12**	0.06	-0.14***	0.08*	1.00

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.6: *Possession of a donor card and reciprocity, using standardized factors from reciprocity items*

	Possession of a donor card		
Standardized factor pos. reciprocity	1.034 (0.123)	0.964 (0.126)	0.879 (0.122)
Standardized factor neg. reciprocity	0.772* (0.084)	0.776* (0.095)	0.747* (0.103)
Risk attitude		0.990 (0.128)	0.919 (0.126)
Agreeableness		0.813 (0.107)	0.845 (0.116)
Openness		1.606** (0.245)	1.433* (0.240)
Conscientiousness		0.827 (0.089)	0.843 (0.101)
Neuroticism		1.040 (0.123)	1.055 (0.142)
Extraversion		1.204 (0.171)	1.073 (0.155)
Altruism		2.468*** (0.610)	2.520*** (0.680)
Advance directive			3.969*** (1.358)
Age			0.954*** (0.009)
Male			0.960 (0.263)
Income per capita in 1000 Euros			1.389* (0.230)
Education			1.225 (0.135)
Ever married			2.066* (0.750)
Constant	0.150*** (0.017)	0.076*** (0.016)	0.150*** (0.068)
Pseudo R^2	0.010	0.080	0.159
χ^2	6.296	44.906	70.604
p-value	0.043	0.000	0.000
Observations	683	683	683

Logit regressions, coefficients displayed as odds ratios; standard errors in parentheses.

Reciprocity and personality variables are standardized.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.7: *General willingness to become an organ donor and reciprocity, using standardized factors from reciprocity items*

	General willingness to become an organ donor		
Standardized factor pos. reciprocity	1.417*** (0.116)	1.359*** (0.120)	1.329** (0.119)
Standardized factor neg. reciprocity	0.772** (0.061)	0.778** (0.068)	0.745** (0.067)
Risk attitude		1.186 (0.107)	1.118 (0.104)
Agreeableness		0.983 (0.091)	1.032 (0.100)
Openness		1.251* (0.121)	1.133 (0.117)
Conscientiousness		0.916 (0.085)	0.932 (0.089)
Neuroticism		1.038 (0.090)	1.068 (0.101)
Extraversion		1.039 (0.097)	1.006 (0.097)
Altruism		1.411* (0.233)	1.638** (0.301)
Advance directive			1.641 (0.415)
Age			0.973*** (0.006)
Male			1.058 (0.187)
Income per capita in 1000 Euros			1.226 (0.153)
Education			1.067 (0.088)
Ever married			0.976 (0.237)
Constant	0.994 (0.079)	0.818 (0.096)	2.030* (0.708)
Pseudo R^2	0.040	0.063	0.098
χ^2	36.602	54.368	75.420
p-value	0.000	0.000	0.000
Observations	683	683	683

Logit regressions, coefficients displayed as odds ratios; standard errors in parentheses.

Reciprocity and personality variables are standardized.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.8: *Possession of a donor card and reciprocity, using combined samples from 2006 and 2009 and reduced reciprocity measures*

	Possession of a donor card		
Positive reciprocity reduced	1.027 (0.109)	0.914 (0.107)	0.870 (0.106)
Negative reciprocity reduced	0.796* (0.080)	0.804 (0.090)	0.808 (0.094)
Risk attitude		1.014 (0.112)	0.939 (0.113)
Agreeableness		0.888 (0.102)	0.892 (0.104)
Openness		1.387* (0.180)	1.200 (0.171)
Conscientiousness		0.816* (0.079)	0.843 (0.087)
Neuroticism		0.980 (0.099)	0.996 (0.111)
Extraversion		1.290* (0.163)	1.318* (0.174)
Altruism		2.475*** (0.548)	2.098** (0.488)
Advance directive			3.083*** (0.831)
Age			0.995 (0.004)
Male			0.744 (0.171)
Income per capita in 1000 Euros			1.134 (0.139)
Education			1.358** (0.129)
Ever married			0.682 (0.173)
Constant	0.140*** (0.014)	0.072*** (0.013)	0.052*** (0.019)
Pseudo R^2	0.007	0.068	0.123
χ^2	5.535	54.532	88.997
p-value	0.063	0.000	0.000
Observations	921	921	921

Logit regressions, coefficients displayed as odds ratios; standard errors in parentheses.

Reciprocity and personality variables are standardized.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.9: *General willingness to become an organ donor and reciprocity, using combined samples from 2006 and 2009 and reduced reciprocity measures*

	General willingness to become an organ donor		
Positive reciprocity reduced	1.495*** (0.110)	1.373*** (0.107)	1.356*** (0.107)
Negative reciprocity reduced	0.824** (0.055)	0.849* (0.061)	0.830* (0.061)
Risk attitude		1.255** (0.096)	1.169* (0.092)
Agreeableness		1.016 (0.080)	1.034 (0.084)
Openness		1.226* (0.102)	1.126 (0.099)
Conscientiousness		0.980 (0.077)	0.990 (0.080)
Neuroticism		0.975 (0.073)	0.986 (0.078)
Extraversion		1.036 (0.084)	1.050 (0.088)
Altruism		1.440** (0.202)	1.482** (0.224)
Advance directive			1.610* (0.353)
Age			0.989*** (0.003)
Male			0.924 (0.137)
Income per capita in 1000 Euros			1.156 (0.121)
Education			1.123 (0.079)
Ever married			0.649* (0.123)
Constant	1.074 (0.073)	0.866 (0.088)	1.249 (0.318)
Pseudo R^2	0.034	0.062	0.090
χ^2	39.601	71.429	94.428
p-value	0.000	0.000	0.000
Observations	921	921	921

Logit regressions, coefficients displayed as odds ratios; standard errors in parentheses.

Reciprocity and personality variables are standardized.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 2

Information Overload and Confirmation Bias

JOINT WITH LORENZ GÖTTE AND BENSON TSZ KIN LEUNG

2.1 Introduction

Confirmation bias refers to the tendency to seek or interpret evidence in ways that affirm one's existing beliefs, expectations, or a hypothesis at hand (Nickerson, 1998). The bias has been well-documented in different contexts, including medical diagnoses (Croskerry, 2003; Pang et al., 2017), judicial decisions (Roach, 2010), financial markets (Farmer, 1999), political polarization (Flaxman et al., 2016; Iyengar and Hahn, 2009) and many others. Understanding its underlying mechanism and driving force is important for improving policies and enhancing social welfare.

Many existing explanations for confirmation bias proposed in the economics literature are preference-related. For instance, Akerlof and Dickens (1982), Köszegi (2003) and Brunnermeier and Parker (2005) show that anticipatory utility or belief-dependent utility leads to confirmation bias; Carrillo and Mariotti (2000) and Bénabou and Tirole (2002) demonstrate that confirmation bias is a remedy for time inconsistent preferences; Crémer (1995) and Aghion and Tirole (1997) explain it with interpersonal strategic concerns. However, we are not aware of direct experimental tests examining which mechanisms lead to belief formation that exhibits confirmation bias.

In this paper, we follow a nascent literature that emphasizes the role of cognitive constraints in giving rise to confirmation bias. A small, but growing theoretical literature analyzes how limited ability/information overload can explain a number of behavioral biases, including confirmation bias and wishful processing (Compte and Postlewaite, 2012; Jehiel and Steiner, 2019; Leung, 2020; Wilson, 2014).

The literature builds on the premises that people are constrained by limited cognitive ability and cannot process all information perfectly. As a result, they have to efficiently allocate their cognitive resources to process different pieces of information which could lead to selective and biased belief updating behavior. This is in particular relevant in the current information era. The internet is a powerful information tool as we can access all imaginable kinds of information with just a fingertip. However, there is way more information than an individual can process given human cognitive constraints. A person could ignore belief-challenging information with the

motive of saving cognitive resources for future signals, but on the other hand could be willing to use her limited cognitive resources to process information which is in line with her existing beliefs. In contrast, in an environment without information overload, he is prone to process both types of information and there is less incentive to practice selective information processing behavior.

To see how a processing constraint could lead to confirmation bias, consider an individual who has to form a belief about an unknown binary state. Suppose he receives two sequential signals with no return, which can either be belief-confirming or belief-challenging (but which are of equal strength). However, the individual can only update her belief with one of the two signals due to cognitive constraints. If the first signal were belief-confirming, updating her belief would yield a posterior that would be difficult to alter by the subsequent signal. The opportunity cost of passing on the second signal is low and thus the individual processes the belief-confirming signal. Conversely, if the first signal were belief-challenging, if the individual updates her belief with it, her posterior would be closer to 50-50, in the sense that there is more uncertainty. In this case, being able to process the second signal would yield a large benefit and thus the individual would rather ignore the belief-challenging signal and save her processing capacity for the second signal. Put differently, belief-confirming and belief-challenging signals lead to a different value of future information which leads to selective and biased information processing behavior.

In our experiment, subjects receive a sequence of numbers as signals about an underlying state of the world, which could be the state “low” or the state “high”. The “low” distribution is more likely to generate small numbers as signals while the “high” distribution is more likely to generate large numbers. Subjects have to navigate through the sequence within 30 seconds and then report their beliefs of the underlying distribution. We impose a stronger information overload in the treatment than in the control condition. An obvious but naive way to vary the level of information overload is to provide the subjects in the treatment group with more signals, however, in that case the effect of information overload would be convoluted with any effects driven by the differences in signals seen by the subjects in the treatment and control condition.

In order to disentangle the effect of information overload, we provide the subjects in the treatment and control group with the same sequence of signals but on different interfaces. In each round of the experiment, two subjects are matched and are assigned to either the treatment or control condition. Importantly, the interface in the treatment condition imposes bigger difficulties for the subjects to process information, which induces a stronger information overload according to the definition in Speier et al. (1999): “Information overload occurs when the amount of input to a system exceeds its processing capacity.” More specifically, in the control condition, as subjects navigate through the sequence, the preceding numbers remain visible and they observe multiple numbers at the same time. In contrast, in the treatment condition, the previous signals disappear so they only see one signal at a time, and thus have to memorize the previous signals.¹ As information is held constant in the treatment and in the control group but belief updating is more cognitively taxing

¹In the control condition, subjects do not have to remember the numbers they have seen, and as they see multiple numbers at the same time, it is easier to develop an idea about the aggregate information conveyed by the numbers.

in the treatment condition, the setup imposes stronger information overload on the subjects in the treatment condition than on the subjects in the control condition.

Building on the theoretical insights, we define confirmation bias as an asymmetric belief updating behavior. We say that a stronger information overload drives a stronger confirmation bias if subjects update more with belief-confirming information and less with belief-challenging information in the treatment condition than in the control condition. Indeed, we find that upon receiving belief-challenging information, subjects in the treatment condition update less compared to the subjects in the control condition. On the other hand, upon receiving a belief-confirming signal, subjects' belief updating behaviors do not differ significantly across the two conditions. Thus, holding the available signals constant, stronger information overload (empirically) leads to more biased processing behavior. The stronger bias is driven by a stronger under-reaction to belief-challenging information, but not by the updating behavior with belief-confirming information. As a result, subjects in the treatment condition are also less likely to switch between guessing "high" and "low" than in the control condition, even when they receive strong belief-challenging signals.

Our findings constitute direct evidence linking a novel mechanism to confirmation bias, and have novel implications. Our findings that information overload makes individuals more prone to confirmation bias suggests that besides preferences, informational environment also contributes to the bias. This novel channel leads to policy implications that are different from the policy implications of utility-based mechanisms. Much like in the experiment reported here, our results imply that one way to weaken confirmation bias is to make information easier to process. Lastly, our findings are particularly pertinent to different social issues in the current information age, with the leading example being ideological polarization (Flaxman et al., 2016; Gentzkow and Shapiro, 2011).

Our paper also contributes to the experimental literature on how individuals update their beliefs. Eil and Rao (2011), Ertac (2011), Grossman and Owens (2012) and Möbius et al. (2014) study how subjects' beliefs about their task performances, IQ or beauty scores evolve with information, and find evidence supporting the phenomenon of overconfidence. Enke and Zimmermann (2019) find that large proportion of their subjects neglect correlation between signals when they form their beliefs. Liang (2019) shows that subjects substantially discount the signals when the quality of information source is ambiguous.

The remainder of this paper is organized as follows. In the next section, we present a simple theoretical model to motivate our hypotheses. In section 3, we outline the experimental design. In section 4, we present descriptive statistics for our sample, as well as our analysis strategy and hypotheses. In section 5, we present the results. Section 6 offers some concluding remarks.

2.2 Model and Intuition

To motivate the intuition that information overload could give rise to confirmation bias, we present a (toy) model following Leung (2020). While the example does not perfectly match our experimental design, it comprises all the key features of our experimental design and illustrates the theoretical foundation of our hypotheses. Consider a subject who has to guess whether the "high" distribution or "low" dis-

tribution was randomly chosen in each round to generate the numbers he observes as signals. If he makes the correct guess, he gets 1 util; otherwise, he gets 0. His prior belief is denoted by $(p_H, 1 - p_H)$ where p_H is the prior probability assigned to the “high” distribution. We assume $p_H > 0.5$.

Before he makes a guess, he receives two signals, denoted by s_1 and s_2 . Each signal is either a high or low number, denoted by h and l respectively. The “high” distribution is more likely to generate a high number while the “low” distribution is more likely to generate a low number. Formally, $s_i = h$ with probability f when the “high” distribution is chosen and correspondingly, with probability $1 - f$ when the “low” distribution is chosen, where $f > p_H > 0.5$.²

We analyze a setting with information overload where the amount of information exceeds the subject’s processing capacity. We assume that the subject forms beliefs by including information from his memory but that he can only memorize one of two numbers he sees. He engages in the following sequential game: After seeing the first number s_1 , he decides whether to spend time and memorize the number. If he chooses to memorize s_1 , he forgoes s_2 ; on the other hand, if he chooses to ignore s_1 , he proceeds to the next signal s_2 and memorizes it. His posterior belief given his memory of a high or a low number, i.e. $m \in \{h, l\}$, is defined as $\tilde{p}_H(m)$, and follows a simple Bayes’ rule:

$$\tilde{p}_H(m) = \begin{cases} \frac{p_H f}{p_H f + (1 - p_H)(1 - f)} & \text{if } m = h; \\ \frac{p_H(1 - f)}{p_H(1 - f) + (1 - p_H)f} & \text{if } m = l; \end{cases} \quad (2.1)$$

The equation implies that the subject naively draws from his memory as if he had only received the one signal that he memorizes. This is however not crucial to the result, and the implication of confirmation bias also holds in a setting where the subject is more sophisticated.³ The proposition characterizes the subject’s optimal decision and contrasts it with a benchmark without information overload. For the simplicity of exposition, we assume the subject memorizes any signal in case of indifference.

Proposition 1. *With information overload, the subject memorizes (ignores) s_1 if it is belief-confirming, i.e., $s_1 = h$ (belief-challenging, i.e., $s_1 = l$). On the other hand, without information overload, i.e., if the subject’s memory capacity is equal to the amount of information (2 in our model), he memorizes both signals.*

Proof. First, we analyze the case where $s_1 = h$. If he memorizes the signal, his posterior favors the “high” distribution and guesses “high”. His utility of memorizing $s_1 = h$ is:

$$\frac{p_H f}{p_H f + (1 - p_H)(1 - f)}$$

² $f > p_H$ ensures the signals are convincing enough such that the subject switches to guessing “low” after he updates his belief with a low number.

³In particular, the result holds when we extend this assumption so that the subject rationally infers information from his selective decision on memorizing a high versus a low number. In that case, the optimal decision is characterized by a Bayesian Nash equilibrium that requires both, optimality of memorizing decision of the subject’s first period self and the consistency of his second period self’s belief. See the online appendix of Leung (2020) for the analysis.

If he ignores the first signal, his guess would depend on the second number, and he guesses “high” if and only if $s_2 = h$. His utility of ignoring $s_1 = h$ is:

$$\underbrace{\left(\frac{p_H f}{p_H f + (1 - p_H)(1 - f)} f + \frac{(1 - p_H)(1 - f)}{p_H f + (1 - p_H)(1 - f)} (1 - f) \right)}_{\Pr(s_2=h|s_1=h)} \underbrace{\frac{p_H f^2}{p_H f^2 + (1 - p_H)(1 - f)^2}}_{\Pr(\text{high distribution}|s_1=s_2=h)} + \underbrace{\left(\frac{p_H f}{p_H f + (1 - p_H)(1 - f)} (1 - f) + \frac{(1 - p_H)(1 - f)}{p_H f + (1 - p_H)(1 - f)} f \right)}_{\Pr(s_2=l|s_1=h)} \underbrace{\frac{(1 - p_H)f(1 - f)}{p_H f(1 - f) + (1 - p_H)f(1 - f)}}_{\Pr(\text{low distribution}|s_1=h, s_2=l)}$$

After some simple algebra, given $s_1 = h$, he memorizes the first number if and only if:

$$p_H f \geq p_H f^2 + (1 - p_H)f(1 - f)$$

which is true as $p_H > 0.5$. Thus he memorizes $s_1 = h$.

Now suppose $s_1 = l$. If he memorizes the signal, he will guess “low” and thus his utility is

$$\frac{(1 - p_H)f}{p_H(1 - f) + (1 - p_H)f}$$

which is closer to 0.5 than if he received a high signal and guessed high, as $p_H > 0.5$. If he ignores the first signal, he guesses “high” if and only if $s_2 = h$. His utility of ignoring $s_1 = l$ is:

$$\left(\frac{p_H(1 - f)}{p_H(1 - f) + (1 - p_H)f} f + \frac{(1 - p_H)f}{p_H(1 - f) + (1 - p_H)f} (1 - f) \right) \frac{p_H f(1 - f)}{p_H f(1 - f) + (1 - p_H)f(1 - f)} + \left(\frac{p_H(1 - f)}{p_H(1 - f) + (1 - p_H)f} (1 - f) + \frac{(1 - p_H)f}{p_H(1 - f) + (1 - p_H)f} f \right) \frac{(1 - p_H)f(1 - f)}{p_H f(1 - f) + (1 - p_H)f(1 - f)}$$

Again, after some simple algebra, given $s_1 = h$, he memorizes the first number if and only if:

$$(1 - p_H)f \geq f(1 - f)$$

which is true as $f > p_H > 0.5$. Thus he memorizes $s_1 = l$. The second part of the proposition is straightforward as the subject incurs no (opportunity) cost of memorizing signals. \square

To understand the intuition of the result, note that the subject trades off between allocating his updating capacity to the current signal s_1 and the future signal s_2 . Roughly speaking, he compares the value of the current and of future information. When s_1 confirms his belief, he becomes more confident that the “high” distribution was drawn. As a result, the value of the future signal s_2 decreases (to 0 in this simple example). In contrast, when s_1 is belief-challenging, the subject’s belief moves towards $(\frac{1}{2}, \frac{1}{2})$. This increases the value of the future information s_2 as he becomes more uncertain about the state. This asymmetry in the value of future information drives confirmation bias, as the subject tends to update his belief with a belief-confirming signal and to stop looking for future information, but to ignore a belief-challenging signal and save his cognitive resources for future information. The result suggests that information overload leads to biased belief updating behavior (see also Jehiel and Steiner, 2019; Wilson, 2014).

2.3 Experimental Design

In the experiment, subjects have to complete 12 rounds of a guessing task, which involves belief-updating with multiple signals.

2.3.1 States and Information of The Guessing Task

The guessing task is designed to investigate how subjects update their beliefs in the face of information overload. In each round of the guessing task, subjects receive two sequences of numbers which are drawn independently from either a “high” or “low” distribution. The set of numbers are integers from 1 to 8, inclusive. The probability distribution of a drawn number given a “high” or “low” distribution is shown in Figure 2.1.

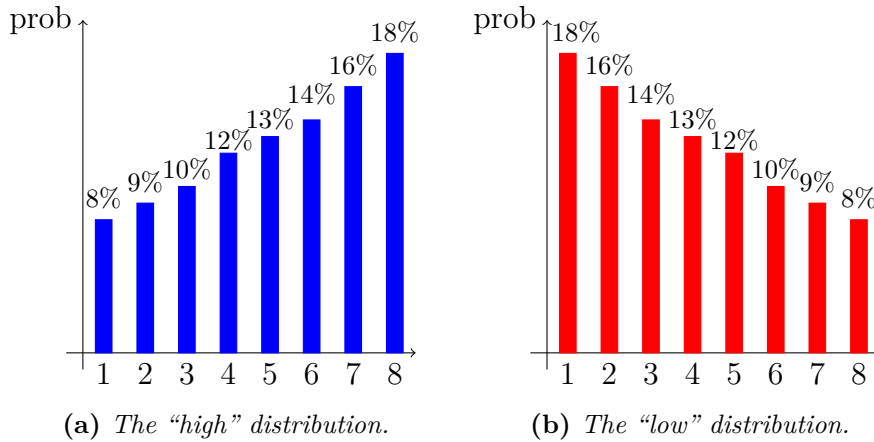


Figure 2.1: *The two distributions shown in bar charts.*

As shown in Figure 2.1, the “high” distribution is more likely to generate larger numbers, while the “low” distribution is more likely to generate smaller numbers. Therefore, subjects can infer which distribution generates the numbers they observe in a particular round. The reasoning behind the parameters of two distributions is explained in detail in Appendix 2.A. Briefly speaking, the two distributions are designed to be sufficiently informative so that subjects could easily make inferences, while not to being too informative to ensure that the probability of receiving belief-challenging information is significant for the analysis. Lastly, the informativeness of a signal should increase steadily as it goes towards the two extremes (numbers 1 and 8).

To make the task more accessible to subjects, we call the distributions that generate the numbers “computers” where the “high” (“low”) computer is more likely to generate high (low) numbers.

2.3.2 The procedure of the guessing task

Pairing and assignment of treatment and control. All subjects play 12 rounds of the guessing task and are assigned to the treatment and control condition alternately. In the beginning of each round, each treatment subject is randomly matched with a control subject to form a pair, and two pairs are randomly matched

Table 2.1: *An example of the underlying distributions and numbers for subjects in two pairs which belong to the same super-pair.*

Super-pair	Pair	Subject	Condition	Underlying distribution	Numbers
Super-pair 1	Pair 1	Subject 1	Treatment	“Low”	2, 3, 1, 4, 5, 3, 6, 1, ...
		Subject 4	Control	“Low”	2, 3, 1, 4, 5, 3, 6, 1, ...
	Pair 2	Subject 2	Treatment	“High”	7, 6, 8, 5, 4, 6, 3, 8, ...
		Subject 3	Control	“High”	7, 6, 8, 5, 4, 6, 3, 8, ...

to form a super-pair.⁴ During the respective round, the two matched subjects in a pair observe the same sequence of numbers drawn from the same underlying distribution, which is either the “high” or “low” distribution with equal probability. This is illustrated in Table 2.1 as subject 1 and 4 (or 2 and 3), who belong to the same pair, see the same numbers drawn from the same underlying distribution. As mentioned in the introduction, it allows us to single out the effect of information overload, by keeping the two subjects’ available information constant.

On the other hand, the two matched pairs in a super-pair see numbers drawn from different underlying distributions, and the numbers received by the two pairs are symmetric around $4\frac{1}{2}$ (and add up to 9). Given the symmetry of the two distributions⁵, the numbers they receive are of the same strength but support different underlying distributions. This is illustrated in Table 2.1. The numbers seen by pairs 1 and 2, which belong to the same super-pair, are symmetric around $4\frac{1}{2}$ and drawn from different distributions. First, comparing the beliefs of subjects in the two matched pairs allows us to test whether there is any intrinsic bias towards either of the two distributions. Second, if there is no bias towards either of the two distributions, the subjects in the two matched pairs should have exactly opposite beliefs, i.e., if subject 1 believes the “low” distribution has been chosen with probability x after seeing a sequence of numbers, subject 2 should believe the “high” distribution has been chosen with probability x . Thus, with careful normalization, this allows us to leverage on the symmetry of the two distributions to increase our statistical power, as it essentially doubles the observations of belief-updating with the same sequence of numbers.

The timeline of a round of the guessing task is illustrated in Figure 2.2 and explained in detail below.

First belief elicitation and belief elicitation mechanism. Before the subjects have seen any numbers, we conduct the first belief elicitation at the beginning of each round. We use a variant of the Becker-DeGroot-Marschak method (Becker et al., 1964), which is shown in Figure 2.3. First, subjects have to guess whether the “high” or “low” distribution has been selected. Second, they have to choose between the following two options: earn 8€ if their chosen distribution is selected, or earn

⁴The numbers of subjects of every sessions are restricted to even numbers, but not to multiples of 4.

⁵The probability of seeing a number x with the “low” distribution is equal to the probability of seeing a number $9 - x$ with the “high” distribution.

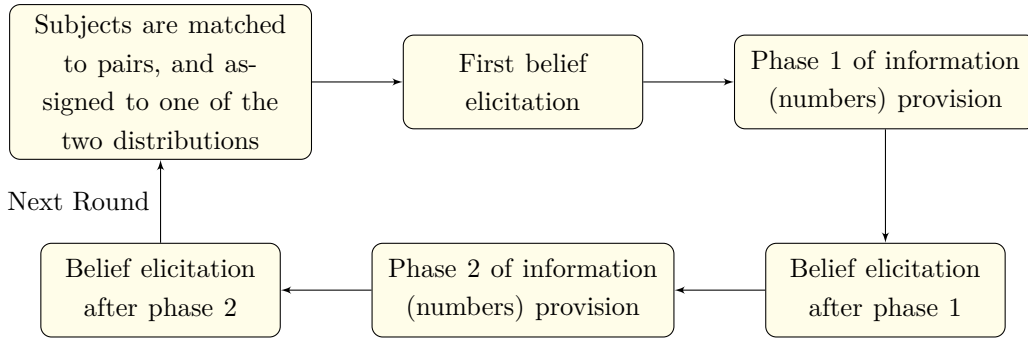


Figure 2.2: *Sequence of a round.*

8€ with probability $x\%$, where x starts at 50% and increases in 5% increments per row. The mechanism is incentive compatible. As an example, if a subject believes the “high” distribution has been chosen with 66%, he should choose “high” for the first question and for the second question he should switch from option 1 to option 2 when $x = 70$ as shown in the figure.

This first belief elicitation is used to ensure that subjects hold the 50-50 belief before seeing any numbers (and that they understand they are at the beginning of a new round). There is a soft time limit of 30 seconds⁶ for the belief elicitation. Afterwards, subjects see two sequences of numbers drawn by the selected distribution in two phases, with a second belief elicitation in between the two phases and a third belief elicitation after the second phase.

Phase 1 of information (numbers) provision. In the first phase, all subjects see 5 numbers displayed on the screen for 30 seconds (Figure 2.4). The two matched subjects from treatment and control condition see the same 5 numbers. After 30 seconds, the subjects are redirected to the page of the second belief elicitation.

Belief elicitation after phase 1. After phase 1, we elicit the subjects’ beliefs using the same table shown in Figure 2.3. Their choices in the first belief elicitation are shown as defaults. The first phase, which shows 5 numbers to the subjects, naturally induces heterogeneous beliefs across all the subjects. Thus, it allows us to define belief-confirming and belief-challenging information and to study how belief updating is different with the two types of information.

Phase 2 of information (numbers) provision. In the second phase, subjects can see up to 7 numbers with a strict time limit of 30 seconds. Paired subjects in the treatment and control conditions see the same numbers but with a different screen layout.

The treatment condition The layout and flow of the treatment condition is illustrated in Figure 2.5. The subjects see one number at a time. They can decide when to advance the sequence by clicking the blue “Next” button. Upon clicking “Next”, the next number in the sequence is revealed and the preceding number disappears. Moreover, the subject is unable to return to the preceding numbers.

⁶If a subject has spent more than 30 seconds, he is shown a warning message which reminds him that time is up. However, subjects are not automatically redirected to the next page.

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

Option 1		Option 2
Win 8€, if "high" is true	<input checked="" type="radio"/> <input type="radio"/>	Win 8€ with probability 50 %
Win 8€, if "high" is true	<input checked="" type="radio"/> <input type="radio"/>	Win 8€ with probability 55 %
Win 8€, if "high" is true	<input checked="" type="radio"/> <input type="radio"/>	Win 8€ with probability 60 %
Win 8€, if "high" is true	<input checked="" type="radio"/> <input type="radio"/>	Win 8€ with probability 65 %
Win 8€, if "high" is true	<input type="radio"/> <input checked="" type="radio"/>	Win 8€ with probability 70 %
Win 8€, if "high" is true	<input type="radio"/> <input checked="" type="radio"/>	Win 8€ with probability 75 %
Win 8€, if "high" is true	<input type="radio"/> <input checked="" type="radio"/>	Win 8€ with probability 80 %
Win 8€, if "high" is true	<input type="radio"/> <input checked="" type="radio"/>	Win 8€ with probability 85 %
Win 8€, if "high" is true	<input type="radio"/> <input checked="" type="radio"/>	Win 8€ with probability 90 %
Win 8€, if "high" is true	<input type="radio"/> <input checked="" type="radio"/>	Win 8€ with probability 95 %

Next

Figure 2.3: *The belief elicitation screen.*

Round 1, Phase 1

Time left for this phase: 0:28

5 7 3 7 3

Figure 2.4: *A screen shot of Phase 1.*

Subjects in the treatment condition face a trade-off between spending more time on the current number and saving time for the next numbers.

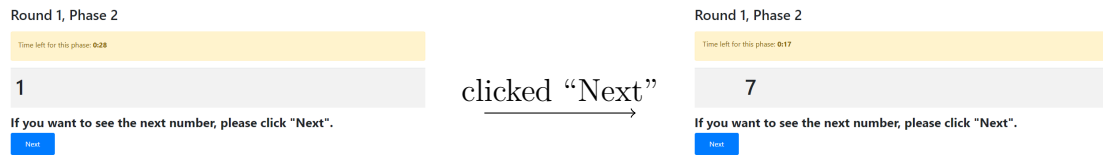


Figure 2.5: *A screen shot of phase 2 in the treatment condition.*

The control condition In contrast, subjects in the control condition cannot influence when the next number appears, while the preceding numbers do not disappear when additional numbers are displayed (Figure 2.6). They start with one number on their screen and as they advance through the sequence, they see two, three, four (etc.) numbers on the screen at the same time. To ensure the information they receive and the timing of information provision are the same as their counterpart in the treatment condition, the control subjects advance in the sequence at the same time as their matched treatment subject click “Next”. The treatment subjects, however, are not aware that they can control the other’s advancement in the sequence, nor do the control subjects know that their advancement is controlled by others.

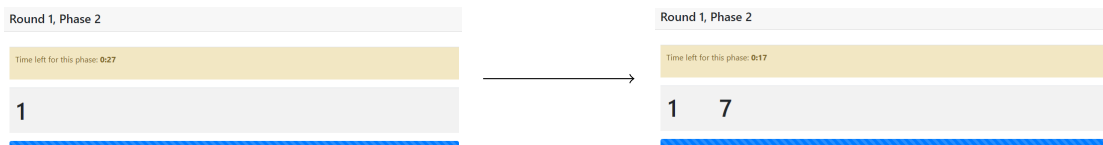


Figure 2.6: *A screen shot of phase 2 in the control condition.*

The main difference between the treatment and control conditions is that it is easier for the control subjects to update their beliefs with all the signals being visible at the same time. In other words, subjects in the control condition are exposed to weaker information overload than those in the treatment condition, despite the fact that they essentially receive the same information.

Belief elicitation after phase 2. After 30 seconds of phase 2, subjects are redirected to the page of the third belief elicitation. Their choices of the second belief elicitation are shown as defaults. By comparing how subjects update their beliefs in the treatment and control condition with belief-confirming and belief-challenging information, we draw insights on how information overload gives rise to confirmation bias.

New round. Following the belief elicitation after phase 2, a new round begins where subjects are randomly re-matched and assigned to either the “high” or “low” distribution with equal probability. Subjects are re-directed to a screen that reminds them that a new round has started.

2.3.3 Procedural Details

We conducted 12 sessions of the experiment, involving 260 subjects in the BonnEconLab at the University of Bonn. The participants were university students and were recruited through the online recruitment system h-root (Bock et al., 2014). The experiment was coded and run in o-Tree (Chen et al., 2016). Each session took about 2h 15min. The subjects were paid according to a randomly drawn decision in the first, second and third belief elicitation (Figure 2.3) from three different randomly chosen rounds. For example, if the first belief elicitation of round 2 is chosen, we randomly choose one of the binary choices in the corresponding belief elicitation table to determine the payment: If the binary choice for the first question “Which Computer is more likely that has been selected?” is chosen, we pay the subject 8€ if the answer is correct; if the binary choice of one of the “option 1 vs. option 2” decision is chosen, we pay according to the option chosen by the subject. We then repeat the same process for the second and third belief elicitation of two other randomly chosen rounds. The maximum earning is thus 24€. The average earnings were 18.12€ per subject, plus a participation fee of 7€.

2.4 Analysis

In this section, we begin by presenting the descriptive statistics for our sample, before constructing and introducing the main variables of interest. Thereafter, we explain our analysis strategy and hypotheses.

2.4.1 Data

Across all 12 sessions, 260 subjects each played 12 rounds of the guessing task. The average age of the subjects was 23, while the maximum age was 30. 109 subjects were male and 151 subjects were female. 50% of the subjects stated that they have taken economics or statistics courses.

Observations

Every subject in each round gives us one observation for the guessing task. Each of the 260 subjects played 12 rounds of the guessing task, which contributed 3120 observations in total. In the first four sessions, unfortunately there was a technical glitch with the computer system recording how many numbers a subject saw in phase 2. More specifically, with some small probability, the recorded number was one fewer than it should be, e.g., the computer system may have recorded 6 while the subject has seen 7 numbers in phase 2. Thus, we dropped the observations in the first four sessions if the recorded numbers of signals seen in phase 2 were less than 7, which amounts to 264 of 1008 observations.⁷ The technical glitch was fixed in later sessions. After dropping the observations as mentioned above, we have 2856 observations in total.

⁷As shown in Table 2.2 in almost all (97%) of the observations, subjects have seen all 7 numbers in phase 2. We therefore are confident that dropping the data does not systematically affect our results.

Table 2.2: *Frequencies and proportion of observations where 4, 5, 6 and 7 numbers have been seen in phase 2.*

Number of signals seen in phase 2	All Sessions		All sessions, with only observations where first elicited belief equals (50%, 50%)		Sessions 5 - 12, with only observations where first elicited belief equals (50%, 50%)	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
4	2	0.07%	2	0.09%	2	0.12%
5	12	0.42%	9	0.40%	9	0.55%
6	49	1.72%	34	1.52%	34	2.07%
7	2793	97.79%	2196	97.99%	1691	97.27%

Notes. Note that in this table we do not include observations in sessions 1 to 4 where the number of signals seen in phase 2 is less than 7, but have included the observations with non-(50%, 50%) first elicited belief.

Furthermore, in the analysis, we only use the observations where choices in the first belief elicitation are compatible with the belief that the underlying distribution is “high” or “low” with equal probability, i.e., they chose option 2 in all rows in the table shown in Figure 2.3, except possibly in the first row. This is a test to ensure that subjects understood the belief elicitation mechanism, and that they were at the beginning of a new round and a new distribution has been drawn with equal probability. In 615 of the 2856 observations (i.e., 21.53%), subjects’ choices in the first belief elicitation fail the test⁸.

For example, in 84 observations, subjects’ choices in the first belief elicitation indicate that they are at least 95% confident about the underlying distribution (that they chose option 1 in all rows in the table shown in Figure 2.3). This might be due to misunderstanding of the belief elicitation mechanism. However, in most cases subjects realized in later rounds that they were filling out the belief elicitation table incorrectly and did not make the same mistake over all 12 rounds. In fact, only 6 subjects chose option 1 for all rows in the first belief elicitation in more than 6 rounds. After excluding observations in which the first elicited belief is not (50%, 50%), we have 2241 final observations. Unless otherwise stated, all our analyses are based on these observations.

Numbers of Signals Seen in Phase 2

Table 2.2 shows the descriptive statistics of the number of signals subjects have seen in phase 2. We see that only a small fraction of subjects have seen less than 7 numbers in phase 2. For instance, in sessions 5 to 12, where the technical glitch mentioned in the previous section was fixed, less than 3% of the subjects saw less than 7 numbers, i.e., almost all subjects managed to reveal all 7 numbers in the 30 seconds time limit. Thus, we are confident that the results presented in this paper are not artificially created by the fact that we have dropped the observations in the first four sessions where the number of signals seen was less than 7.

⁸Among the subjects who have never taken any statistics or economics courses, 26.70% of the observations exhibit choices in the first belief elicitation that are not compatible with belief of equal probability on the two distributions, while the proportion is 16.27% for the subjects which have taken statistics or economics courses.

Randomization of High and Low States

Among all 2241 final observations, 48.77% are assigned to the “high” distribution while 51.23% are assigned to the “low” distribution. The composition is not exactly half-half because we have dropped some observations as mentioned before.

Treatment and Control Condition

Each subject was alternately assigned to the treatment and control conditions in the 12 rounds of the guessing task. Among the final 2241 observations, 1124 observations are from the control condition and 1117 observations are from the treatment condition. Again, the composition is not exactly half-half because we have dropped some observations as mentioned before.

2.4.2 Variables of Interest

Elicited Beliefs

The first elicited belief is denoted by p_0 , while the elicited beliefs after phase 1 and phase 2 are denoted by p_1 and p_2 respectively. All elicited beliefs are normalized such that p_0 , p_1 and p_2 represent the (subjective) probability of the “high” distribution being chosen. Note that $p_0 = 0.5$ in all observations used in the analysis, as we exclude those whose first elicited belief does not equal to 0.5. On the other hand, p_1 and p_2 are pinned down by the point subjects switch from option 1 to option 2 in the belief elicitation table. As an example, if in the belief elicitation after phase 1, subjects guess “high” for the first question as shown in Figure 2.3, and switch from option 1 to option 2 when the winning probability of the random lottery is 70%, we define $p_1 = 0.675$, i.e., the average belief compatible with those choices. In the analysis, p_1 , is treated as the prior belief of the subjects, and we investigate whether subjects update their beliefs upon receiving the belief-confirming or belief-challenging signals in phase 2 differently in the control and treatment conditions.

Bayesian Beliefs

The Bayesian counterpart of the first elicited belief is denoted by p_0^B and correspondingly, the Bayesian counterparts of the elicited beliefs after phase 1 and phase 2 are denoted by p_1^B and p_2^B respectively.

p_0^B is always equal to 0.5 as the distribution is drawn with equal probability. The Bayesian belief after phase 1, p_1^B , is constructed using the first elicited belief p_0 .

$$\frac{p_1^B}{1 - p_1^B} = \prod_{s_i \in \mathcal{S}_1} \frac{f_H(s_i)}{f_L(s_i)} \times \frac{p_0}{1 - p_0}, \quad (2.2)$$

where \mathcal{S}_1 denotes the set of the 5 numbers in phase 1, and $f_H(s_i)$ is the probability that the “high” distribution generates number s_i while $f_L(s_i)$ is the probability that the number is drawn from the “low” distribution. Since we only include observations where $p_0 = 0.5$, this equals to

$$\frac{p_1^B}{1 - p_1^B} = \prod_{s_i \in \mathcal{S}_1} \frac{f_H(s_i)}{f_L(s_i)} \times \underbrace{\frac{0.5}{1 - 0.5}}_{=1}.$$

Similarly, the Bayesian belief after phase 2, p_2^B , is constructed using the elicited belief after phase 1, p_1 :

$$\frac{p_2^B}{1 - p_2^B} = \prod_{s_i \in \mathcal{S}_2} \frac{f_H(s_i)}{f_L(s_i)} \times \frac{p_1}{1 - p_1}, \quad (2.3)$$

where \mathcal{S}_2 is the set of numbers seen in phase 2. In other words, p_2^B is equal to the belief of a Bayesian individual if he takes his prior belief p_1 as given⁹ and updates his belief with \mathcal{S}_2 in a statistically optimal way.

Treatment and control condition

For each observation i , the condition imposed on the subject is denoted by the dummy variable T_i , which takes on the value of 1 if the subject is assigned to the treatment condition, and 0 otherwise.

2.4.3 Empirical Strategy and Hypothesis

To examine how information overload plays a role in confirmation bias, we analyze two indicators of confirmation bias, namely switching behavior and changes in belief.

Switching Behavior

The first indicator we analyze pertains to the switching decisions of the subjects. A switch is defined as the scenario where a subject guessed “high” after phase 1 but guessed “low” after phase 2, or vice versa. Moreover, we say that a subject has made a switching mistake when his switching decision is different from that of a Bayesian individual. We analyze two different switching mistakes. The first mistake is the case where the subjects should switch if they were Bayesian but they ended up not switching; the second mistake is the case where the subjects should not switch if they were Bayesian but they ended up switching.

If information overload induces a stronger confirmation bias, subjects in the treatment condition should update (weakly) more to belief-confirming information and conversely, update (weakly) less to belief-challenging information relative to their counterpart in the control condition. Thus, they should be less likely to switch their decisions, which leads us to the following hypotheses:

Hypothesis 1W. *Subjects in the treatment condition are **weakly** less likely to switch decisions when they should, than their counterparts in the control condition.*

Hypothesis 2W. *Subjects in the treatment condition are **weakly** less likely to switch decisions when they should not, than their counterparts in the control condition.*

The strong form of hypotheses 1W and 2W are as follows:

⁹We believe that it is very unlikely that subjects would revise their belief with respect to the signals in phase 1 during phase 2 given the short time limit and cognitive overload. Furthermore, previous elicited belief is always shown as a default in the next belief elicitation, which provides an anchor to distinguish the belief formation in different phases.

Hypothesis 1S. *Subjects in the treatment condition are **strictly** less likely to switch decisions when they should, than their counterparts in the control condition.*

Hypothesis 2S. *Subjects in the treatment condition are **strictly** less likely to switch decisions when they should not, than their counterparts in the control condition.*

Given the theoretical insights from Leung (2020), as illustrated in our toy model, we expect both hypotheses 1W and 2W to hold and at least one of the two strong hypotheses 1S and 2S to hold¹⁰.

It is worth noting that a subject should switch when he receives belief-challenging signals of sufficient strength, while he should not switch when he receives belief-confirming signals or weak belief-challenging signals. Thus hypothesis 1W/1S and hypothesis 2W/2S corresponds to different scenarios: the former examines the subjects' belief updating behavior with strong belief-challenging signals while the latter analyzes the subjects' belief updating behavior with belief-confirming or weak belief-challenging signals.

Next, we present the regression specifications for our analysis. The notation is as follows: i denotes the observation while $m(i)$ denotes the pair that observation i belongs to. As mentioned before, T_i indicates whether observation i is assigned to the treatment or control condition. $\alpha_{m(i)}$ is the fixed effect for pair $m(i)$ that observation i belongs to, so as to account for the numbers seen by each pair in phases 1 and 2. Furthermore, as we have multiple observations per subject since they play 12 rounds of the guessing task, we cluster standard errors at the subject level. Lastly, we denote $\text{Switch}_i = 1$ if the subject switched decisions in observation i , and $\text{Switch}_i = 0$ otherwise. For hypothesis 1W/1S, we estimate the following regression for all observations i where a theoretical Bayesian subject should switch, i.e., where $(p_2^B - 0.5)(p_1 - 0.5) < 0$:

$$1 - \text{Switch}_i = \beta_0 + \beta_1 T_i + \alpha_{m(i)} + \epsilon_i. \quad (2.4)$$

β_1 measures the treatment effect on switching mistakes (not switching when the subject should switch), and hypothesis 1W (1S) translates to $\beta_1 \geq (>)0$. Similarly, for hypothesis 2W/2S, we estimate the follow regression for all observations i where a theoretical Bayesian subject should not switch:

$$\text{Switch}_i = \beta_0 + \beta_1 T_i + \alpha_{m(i)} + \epsilon_i, \quad (2.5)$$

and similarly hypothesis 2W (2S) translates to $\beta_1 \leq (<)0$.

Quantifying Bias

For the second indicator, we quantify subjects' biases in belief formation. We proceed by drawing an analogy between the evolution of the elicited belief to the Bayesian formula.

Consider a subject whose elicited belief after phase 1 is equal to p_1 . After he has seen n numbers in phase 2, in which the set is denoted as \mathcal{S}_2 , his Bayesian belief after phase 2, p_2^B , is given by:

$$\frac{p_2^B}{1 - p_2^B} = \prod_{s_i \in \mathcal{S}_2} \frac{f_H(s_i)}{f_L(s_i)} \times \frac{p_1}{1 - p_1}, \quad (2.6)$$

¹⁰Note that if both hypotheses 1W and 2W hold, while both strong hypotheses 1S and 2S do not hold, subjects' switching decisions do not differ significantly in treatment and control condition.

where $f_H(s_i)$ and $f_L(s_i)$ are the probabilities of seeing numbers s_i when the “high” or “low” distribution is chosen respectively. The product of the odds ratios $\prod_{s_i \in \mathcal{S}_2} \frac{f_H(s_i)}{f_L(s_i)}$ measures the relative likelihood of seeing the numbers in \mathcal{S}_2 with the “high” distribution over that with the “low” distribution. For simplicity of notation, we denote $\prod_{s_i \in \mathcal{S}_2} \frac{f_H(s_i)}{f_L(s_i)}$ by y_{obj} , or as the “objective odds ratio”. Note that the objective odds ratio is a sufficient statistic for a Bayesian individual to update his belief.

We now use the elicited beliefs after phase 1 (p_1) and the elicited beliefs after phase 2 (p_2) to characterize the subjective counterpart of the objective odds ratio, which is denoted as y_{sub} :

$$\begin{aligned} \frac{p_2}{1-p_2} &= y_{sub} \times \frac{p_1}{1-p_1} \\ y_{sub} &= \frac{p_1}{1-p_1} \times \frac{1-p_2}{p_2}. \end{aligned} \tag{2.7}$$

y_{sub} measures the subject’s perceived relative likelihood of seeing the numbers in \mathcal{S}_2 with the “high” distribution over that with the “low” distribution. When $y_{sub} > y_{obj}$, the perception of the subject is biased towards the “high” distribution; when $y_{sub} < y_{obj}$, the perception of the subject is biased towards the “low” distribution.

As mentioned before, if the treatment condition induces a stronger confirmation bias, the treatment subjects update (weakly) more to belief-confirming information but update (weakly) less to belief-challenging information than subjects in the control condition. We denote the subjective odds ratio of the subjects in the treatment and control condition by y_{sub}^T and y_{sub}^C respectively, such that $y_{sub}^T > y_{sub}^C$ implies that subjects are more biased towards the “high” distribution in the treatment condition than in the control condition. We have the following hypotheses:

Hypothesis 3W. *Suppose the numbers seen by the subjects in phase 2 are belief-challenging on aggregate, i.e., $(p_1 - 0.5)(y_{obj} - 1) < 0$. The subjective odds ratio of the subject in the treatment condition is **weakly** more biased towards his prior belief than that of his matched subject in the control condition, i.e., $(p_1 - 0.5)(y_{sub}^T - y_{sub}^C) \geq 0$.*

Hypothesis 4W. *Suppose the numbers seen by the subjects in phase 2 are belief-confirming on aggregate, i.e., $(p_1 - 0.5)(y_{obj} - 1) > 0$. The subjective odds ratio of the subject in the treatment condition is **weakly** more biased towards his prior belief than that of his matched subject in the control condition, i.e., $(p_1 - 0.5)(y_{sub}^T - y_{sub}^C) \geq 0$.*

The strong form of the hypotheses 3W and 4W are as follows:

Hypothesis 3S. *Suppose the numbers seen by the subjects in phase 2 are belief-challenging on aggregate, i.e., $(p_1 - 0.5)(y_{obj} - 1) < 0$. The subjective odds ratio of the subject in the treatment condition is **strictly** more biased towards his prior belief than that of his matched subject in the control condition, i.e., $(p_1 - 0.5)(y_{sub}^T - y_{sub}^C) > 0$.*

Hypothesis 4S. *Suppose the numbers seen by the subjects in phase 2 are belief-confirming on aggregate, i.e., $(p_1 - 0.5)(y_{obj} - 1) > 0$. The subjective odds ratio of the subject in the treatment condition is **strictly** more biased towards his prior belief than that of his matched subject in the control condition, i.e., $(p_1 - 0.5)(y_{sub}^T - y_{sub}^C) > 0$.*

Similar to the analysis on switching mistakes, we expect both hypotheses 3W and 4W to hold and at least one of the two strong hypotheses 3S and 4S to hold. For the analysis, we assume a multiplicative relationship between y_{sub} and y_{obj} such that their logarithmic forms follow an additive relationship¹¹. Put differently, we estimate the treatment effect on $\frac{y_{sub}}{y_{obj}}$. Note that with a multiplicative instead of an additive model that would otherwise estimate the treatment effect on $y_{sub} - y_{obj}$, we can interpret the multiplicative constant as an attention weight on the objective odds ratio (see Jehiel and Steiner (2019) for the theory). Moreover, the estimated y_{sub} is always larger than 0.

Our notation is the same as in the analysis for switching behavior: i denotes the observation and $m(i)$ denotes the pair that observation i belongs to. $y_{i,sub}$ and $y_{i,obj}$ denote the subjective and objective odds ratio of observation i respectively. Again, we include pairwise fixed effects $\alpha_{m(i)}$ and cluster standard errors on subject-level. We estimate the following regression for all observations.

$$\log(y_{i,sub}) - \log(y_{i,obj}) = \beta_0 + \beta_1 T_i + \alpha_{m(i)} + \epsilon_i. \quad (2.8)$$

The treatment effect β_1 is interpreted as follows: As the regression is run in logarithmic form, the subject’s subjective odds ratio in the treatment condition is $\exp(\beta_1)$ times that of his matched subject in the control condition, i.e., $y_{sub}^T = \exp(\beta_1) \times y_{sub}^C$. When $\beta_1 > 0$, we have $\exp(\beta_1) > 1$ which means the treatment subject’s subjective odds ratio is larger than that of his matched control subject, and the treatment subject is biased towards the “high” distribution. In other words, $(p_1 - 0.5)(y_{sub}^T - y_{sub}^C) \geq 0$ if and only if $\beta_1(p_1 - 0.5) \geq 0$, and the testing of the hypotheses boils down to a testing of the sign of β_1 .

2.5 Results

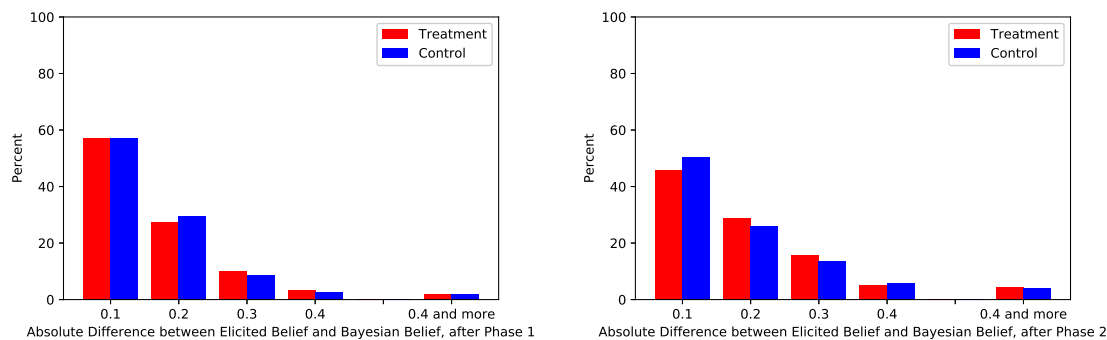
2.5.1 Preliminaries

Before we present the main results for our two indicators of confirmation bias, we first analyze the relationship between elicited belief and Bayesian belief. While it is not the main focus of this paper, the results in this subsection give us a rough idea of how “Bayesian” subjects behave. More importantly, we can ascertain whether subjects understand the experiment well and whether our belief-elicitation mechanism works well in eliciting “normal” behavior of belief-updating.

Figure 2.7a shows the histograms of the absolute difference between elicited and Bayesian belief (the updating mistakes) after phase 1 in the treatment and control conditions, while Figure 2.7b shows the corresponding histograms for beliefs after phase 2. The two graphs show that most of the mistakes (40% – 60%) are less than 10%. Moreover, the frequencies of the mistakes decrease with the magnitude. For example, for belief formation in phase 1 in both treatment and control condition, almost 60% of the elicited beliefs are within 10% difference of the Bayesian beliefs, while only around 10% of the mistakes are as big as 20%.

On the other hand, by comparing the difference between elicited belief and Bayesian belief in the treatment condition after phase 1 to the ones after phase 2 (and control condition after phase 1 and phase 2 respectively), we can see that the

¹¹ $y_{sub} = \lambda y_{obj}$ implies that $\log(y_{i,sub}) = \log \lambda + \log(y_{i,obj})$.



(a) Beliefs after phase 1, treatment versus control. (b) Beliefs after phase 2, treatment versus control.

Figure 2.7: Distribution of the absolute difference between elicited and Bayesian belief.

mistakes in belief formation in phase 2 are in general bigger than the mistakes in phase 1 because of the stronger information overload, i.e., there are more numbers to be processed in the same period of time. For similar reasons, when looking at Figure 2.7b, the mistakes in the belief formation in phase 2 in the treatment condition are in general bigger than in the control condition, i.e., in the control condition, the share of small mistakes is higher than in the treatment condition, while the share of big mistakes is smaller.

Lastly, as we can see in Figure 2.7a, being in the treatment or control condition has no effect on the mistakes made in phase 1, as there are no differences in the settings in phase 1. This is also confirmed by the second and the third column of Table 2.3, which shows that there is no treatment effect on the relationship between elicited and Bayesian belief in phase 1 and thus, no inherent difference between treatment and control condition.

Table 2.3: Analysis of the (absence) of treatment effects after Phase 1, OLS.

	(1) Elicited belief after phase 1	(2) Elicited belief after phase 1	(3) Abs. distance elicited and Bayesian belief after phase 1
Bayesian Belief after phase 1	0.752*** (0.016)	0.752*** (0.016)	
Treatment		-0.00241 (0.006)	0.00153 (0.004)
Constant	0.152*** (0.009)	0.153*** (0.009)	0.109*** (0.003)
R^2	0.701	0.701	0.0000546
Observations	2241	2241	2856
Subjects	235	235	260

Clustered standard errors on subject-level in parentheses.

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

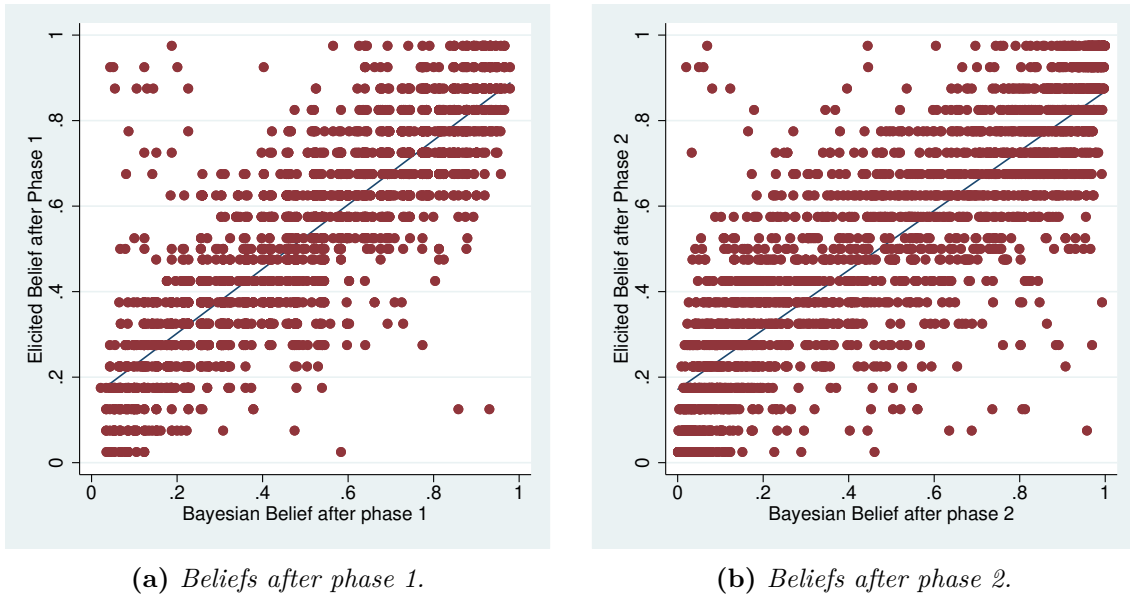


Figure 2.8: Scatter plot and regression line with Bayesian belief on x -axis and Elicited belief on y -axis.

Figures 2.8a and 2.8b show the scatter plots and simple regression lines of elicited beliefs against Bayesian beliefs, after phase 1 and phase 2 respectively. From both figures, we can see that there is a significant and positive correlation between elicited belief and Bayesian belief, which means that subjects understand the essence of the information structure, i.e., higher numbers serve as stronger evidence that the “high” distribution was chosen in the respective round. On the other hand, the slope of the regression line is smaller than 1. Taken together, both findings suggest that on average, subjects believe more in the “high” (“low”) distribution when they receive higher (lower) numbers, but they tend to under-react to signals compared to the Bayesian benchmark. This result coheres with the findings presented in Eil and Rao (2011) and Liang (2019).

2.5.2 Switching Behavior

In this subsection, we analyze the switching behavior of the subjects. Table 2.4 shows the proportion of observations in which the subject has made a switching mistake, in the treatment and the control condition. Note that in the table, we include only complete pairs, i.e., where both subjects in the pair have a first elicited belief that equals 0.5; furthermore, we only include pairs which have the same Bayesian switching choice (e.g., both of them guess “high” after phase 1 and should switch to “low” after phase 2). In total, there are 701 complete pairs with the same Bayesian switching choice.

The first column of the table shows the case where the subjects should switch if they were Bayesian but they ended up not switching. We see that around 36.8% of subjects in the treatment condition did not switch even if they should, while only 27.6% of subjects made such a mistake in the control condition.¹² On the other hand, the second column shows the case where the subjects should not switch but ended up switching. In the treatment condition, subjects are (marginally) less

¹²The numbers are 38.1% vs. 32.6% if we also include incomplete pairs.

Table 2.4: *Proportion of observations in which subjects have made a switching mistake. Only complete pairs with the same Bayesian switching choice are included.*

	Should switch but DID NOT	Should NOT switch but did
Treatment	56/152 $\approx 36.8\%$	30/549 $\approx 5.5\%$
Control	42/152 $\approx 27.6\%$	32/549 $\approx 5.8\%$

likely to switch when they should not than in the control condition.¹³ However, the difference is much smaller compared to the first column, i.e., the difference is 0.4% in the case where the subjects should not switch, while it is 9.2% in the case where the subjects should switch. In both cases, subjects are less likely to switch when exposed to a stronger information overload.

To explore further, we run an OLS regression with pairwise fixed effects, using clustered standard errors at the subject level. As shown in the first column of Table 2.5, the treatment effect is positive and highly significant ($p < 0.01$) in the scenario where the subject should switch but did not, which confirms hypothesis 1S. When there is strong enough belief-challenging information so that the subjects should switch, subjects in the treatment condition are 9.21% less likely to do so than subjects in the control condition.

Table 2.5: *OLS of Switching Decisions after Phase 2.*

	(1) Should switch but didn't	(2) Shouldn't switch but did
Treatment	0.0921*** (0.033)	-0.00364 (0.008)
Constant	0.305*** (0.019)	0.0612*** (0.005)
R^2	0.0280	0.0002
Observations	592	1649
Subjects	207	233

Clustered standard errors on subject-level in parentheses.

Pairwise fixed effects.

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

On the other hand, in the scenario where the subject should not switch but ended up switching, the treatment effect is not significant, which confirms hypothesis 2W, but not hypothesis 2S. In the case where there is no strong enough belief-challenging

¹³If we include the incomplete pairs, the numbers are 5.6% in the treatment condition and 6.5% in the control condition.

information such that subjects should stick to their prior belief, the magnitude of information overload has no effect on switching behavior.

Combining the two findings, we can see that information overload has an asymmetric effect on individuals' switching decision when they receive different types of information. More specifically, the effect is significant only when the subjects should have switched, i.e., they receive strong enough belief-challenging information. Subjects react less to belief-challenging information when they are exposed to stronger information overload, while their reaction to belief-confirming information (that does not induce switching) is unaffected by information overload. This finding suggests that a stronger information overload inhibits switching through individuals' under-reaction to belief-challenging information.

The asymmetric effect also speaks against another possible hypothesis that subjects switch less under stronger information overload only because they under-react to every signal they receive instead of being more biased. First note that the scenarios where they should not switch is predominately composed of cases where they receive belief-confirming information. Suppose in contrast the subjects under-react more to both belief-confirming and belief-challenging information in the treatment condition, they then will be more reluctant to update their beliefs towards the extreme when they receive belief-confirming information. It implies that there should be a higher probability of switching in the treatment condition, and this is clearly rejected by the second column of Table 2.5. The results above are also consistent with the analysis of quantifying bias as will be shown in the following subsection.

2.5.3 Quantifying Bias

To further illustrate the asymmetric effect of information overload on belief updating, we now present the regression analysis of the quantified bias. The quantified bias is represented by $\log(y_{i,sub}) - \log(y_{i,obj})$ as shown in equation (2.8). It measures the direction and magnitude of the discrepancy between the subjective belief updating of the subjects and the Bayesian benchmark.

We first look into the scenario where the numbers seen in phase 2 are in aggregate belief-challenging, i.e., $(p_1 - 0.5)(y_{sub} - 1) < 0$. The results are presented in Table 2.6. The first and second column shows the case where subjects guessed "high" and "low" after phase 1 respectively. In the third column, we pool the two cases by taking advantage of the symmetry of the information structure¹⁴, and this allows us to increase statistical power.

We observe that the treatment effects are significant in all three cases when subjects receive in aggregate belief-challenging information in phase 2. For example in the first column, we see that $\beta_1 = 0.17 > 0$ ($p < 0.05$) such that the subjective odds ratio is $\exp(0.17) = 1.19$ times higher in the treatment condition than in the control condition. This implies that a subject with a "high" prior under-reacts more to belief-challenging information when facing a stronger information overload. Similar conclusions can be drawn from the second and third column. For subjects with "low" priors, we find that $\beta_1 = -0.155$ which is also significant ($p < 0.05$), such

¹⁴We pool the two cases as follows: in the case where subject guessed "low" after phase 1, we normalize the belief as the probability that the "low" distribution is drawn. Odds ratios are also normalized accordingly. Thus a larger belief implies that the subject is more confident about his guess, while a larger odds ratio implies that the signals are "more" belief-confirming.

that the subjective odds ratio is $\exp(-0.155) = 0.856$ times lower in the treatment condition than in the control condition. For the pooled sample, we find that $\beta_1 = 0.164$ ($p < 0.01$) such that the subjective odds ratio is $\exp(0.164) = 1.178$ times higher in the treatment condition than in the control condition. These results all show that subjects react less to belief-challenging information when they are imposed with stronger information overload. Thus, we conclude that the results confirm hypothesis 3S.

Table 2.6: *OLS on quantified bias when numbers seen in phase 2 are in aggregate belief-challenging.*

	(1) High prior, should update downwards	(2) Low prior, should update upwards	(3) Pooled, challenging info
Treatment	0.170** (0.068)	-0.155** (0.073)	0.164*** (0.050)
Constant	0.378*** (0.041)	-0.211*** (0.045)	0.305*** (0.032)
R^2	0.0174	0.0161	0.0169
Observations	516	398	914
Subjects	205	188	225

Clustered standard errors on subject-level in parentheses.

Pairwise fixed effects.

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

Table 2.7: *OLS on quantified bias when numbers seen in phase 2 are in aggregate belief-confirming.*

	(1) High prior, should update upwards	(2) Low prior, should update downwards	(3) Pooled, confirming info
Treatment	-0.0278 (0.057)	0.0552 (0.052)	-0.0411 (0.037)
Constant	-0.774*** (0.035)	0.863*** (0.032)	-0.816*** (0.024)
R^2	0.0005	0.0025	0.0012
Observations	703	624	1327
Subjects	214	211	230

Clustered standard errors on subject-level in parentheses.

Pairwise fixed effects.

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

The results for the second scenario, where numbers seen in phase 2 are on aggregate belief-confirming, are shown in Table 2.7. In contrast with the results for

belief-challenging information, we can see that the treatment effects are not significant in all three cases, even when we pool the subjects with “high” and “low” priors and take advantage of the larger sample size. Thus, the results confirm hypothesis 4W, but not hypothesis 4S.

Combining the results in both tables, we conclude that stronger information overload in the treatment condition induces a stronger confirmation bias which is similar to the analysis of switching behavior: subjects under-react more to belief-challenging information in the treatment condition than in the control condition, while the updating behavior with belief-confirming information is not affected by the magnitude of information overload. The stronger confirmation bias is driven via the under-reaction to belief-challenging information, but not via the updating behavior with belief-confirming information. Similar to the analysis of switching behavior, this asymmetry stands in contrast to the hypothesis that stronger information overload would induce more under-reaction to both belief-confirming and belief-challenging information.

2.6 Conclusion

In this study, we investigate the role of information overload in giving rise to confirmation bias. We show that when subjects are exposed to stronger information overload, their belief updating behavior exhibits a stronger confirmation bias, holding constant the signals they receive. The effect is driven by the increased under-reaction to belief-challenging information while the updating behavior concerning belief-confirming information is unaffected. In addition to the popular view that confirmation bias is driven by intrinsic preferences for belief-confirming information, our findings demonstrate that the bias also strongly depends on the informational environment. This lends credence to the growing theoretical literature which details that limited attention and cognitive ability could explain a number of behavioral anomalies.

This additional channel of confirmation bias has important implications. First, it sheds light on the debate of whether the Internet strengthens biased behavior and promotes ideological polarization. Our results suggest that information overload, as driven by the Internet, could pose substantial problems by driving individuals to ignore belief-challenging information. Thus, the Internet could promote polarization even though it provides more and on average, better information to the public. Simply providing more information might not be a good way to mitigate confirmation bias and the extent of polarization. In particular, this paper suggests that a better solution could be to make it less cognitively demanding to process information.

On the other hand, the results imply that research and policy evaluations have to take into account that confirmation bias — or more generally: information processing behavior — interacts with the informational environment. This effect is absent if one assumes that confirmation bias is solely driven by intrinsic preferences. For example, a mandate for firms to provide more information to consumers may seem welfare-improving. However, such a policy intervention could lead to information overload and exacerbates confirmation bias, which in turn reduces market competition. Ignoring this indirect effect might yield dramatically different results.

2.A Appendix A. The Distributions in the Guessing Task

We present the reasoning behind the parameters of the distributions. In particular, they are chosen to satisfy the following criteria:

1. The logarithmic odds ratios are monotonic and approximately linear, as shown in figure 2.9. That is, higher numbers are a stronger evidence that the “high” distribution is true and the differences in the strengths of adjacent signals are approximately constant;
2. After seeing the first sequence of numbers, there are enough subjects with confident beliefs, i.e., they believe that the state is high (low) with probability 75% or above. Table 2.9 shows that more than 40% of the subjects are “confident” after seeing 5 signals. This is to ensure that there exists a significant number of confident individuals such that confirmation bias could take effect;
3. After seeing the first sequence of numbers, there should not be many subjects with too confident beliefs, e.g., subjects that believe that the state is high(low) with probability 95% or above. Table 2.10 shows that less than 2% of the subjects are extremely confident after seeing 5 signals. This is because the belief elicitation is restricted to increments of 5%. When a subject believes that the state is high with 95% certainty, even if he receives several “number 8”s, the change in his belief is bounded by +5% and is not measurable. Moreover, it ensures that there is a sufficient number of observations where a switching occurs, as shown in table 2.8.

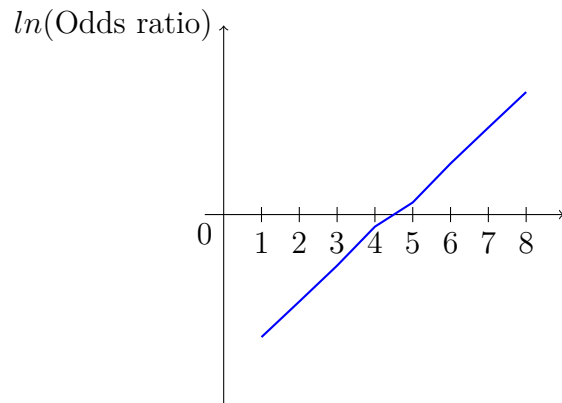


Figure 2.9: *Logarithmic Odds ratios of the numbers 1-8*

min	mean	max
18.5%	24.2%	29.5%

Table 2.8: *Simulated proportion of observations where subjects should switch from believing “High” after phase 1 to believing “Low” after phase 2, or from believing “Low” to believing “High”, with 10,000 simulations of 2,000 observations*

		Belief				
		= 0.5	> 0.6 or < 0.4	> 0.65 or < 0.35	> 0.7 or < 0.3	> 0.75 or < 0.25
5 draws	Average proportion of subjects	0	76.1%	64.7%	52.5%	41.7%
12 draws	Average proportion of subjects	0%	94.4%	88.8%	76.0%	68.9%

Table 2.9: *Distribution of Bayesian beliefs given 5 and 12 draws, with 10,000 simulations of 2,000 observations.*

		Belief			
		> 0.8 or < 0.2	> 0.85 or < 0.15	> 0.9 or < 0.1	> 0.95 or < 0.05
5 draws	Average proportion of subjects	32.2%	19.2%	10.1%	1.65%
12 draws	Average proportion of subjects	60.9%	51.35%	40.05%	24.35%

Table 2.10: *Distribution of Bayesian beliefs given 5 and 12 draws, with 10,000 simulations of 2,000 observations.*

2.B Appendix B: Instructions

Instructions for the first part of the experiment¹

In this part of the experiment you will go through 12 rounds of a task which will be explained to you in the following. Each round will take about 2 - 3 minutes. In this part of the experiment you can win up to 24 €.

1 What is the experiment about?

This part of the experiment consists of 12 rounds of a task which will be described in the following. There are two computers, a "high" computer and a "low" computer: One of these computers, the "high" computer, generates high numbers more frequently, while the "low" computer generates low numbers more frequently. At the beginning of each round, one of the two computers (high or low) is randomly selected, but you do not know which one. The probability for each computer is equal, i.e. 50% for each computer. In each round, you will see numbers which have been generated by the selected computer. We will ask you to indicate your guess which one of the two computers has been selected in this round using the numbers you have seen as indicators for the selected computer.

As you can see in figure 5, we will ask you three times per round to indicate your guess.

- The first time, we will ask you at the beginning of a round, before you have seen any numbers, without any additional information.
- The second time, after you have seen 5 numbers in phase 1, which have been generated by the selected computer of this round.
- The third time, after you have seen additional numbers in phase 2, which have been generated by the selected computer of this round.

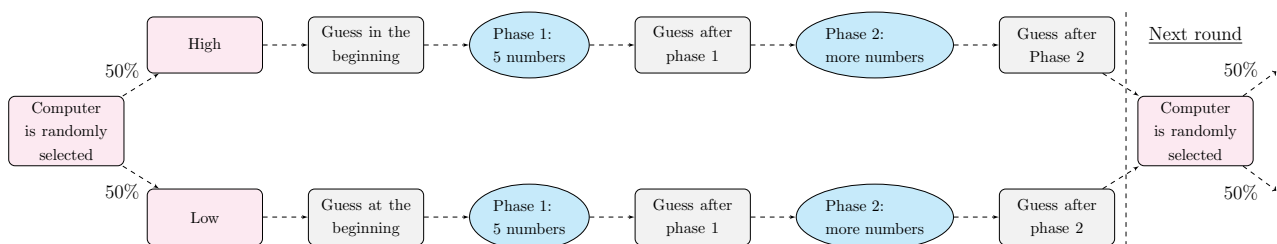


Figure 1: Sequence of a round

¹These instructions were originally in German and have been translated to English. The original German version is available on request.

2 How to make a guess?

In the following, we explain how you can use numbers as indicators for the computer which has been selected at the beginning of a round and how we will retrieve your guess on the monitor.

2.1 The high and the low Computer

Both computer can only generate numbers between 1 and 8. Table 1 shows the probabilities with which the computers produce the numbers 1 to 8. For example, the probability that the “high” computer generates the number 8 is 18%, this means that it happens in 18 out of 100 cases on average. The “high” computer generates smaller numbers less likely. For example, the probability that the computer generates the number 1 is only 8%, in other words, in 8 out of 100 cases.

The “low” computer generates numbers with the probabilities shown in table 2 and can be seen as a mirror image of the “high” computer. For example, the probability with which the “low” computer generates the number 8 is only 8%, on other words, in 8 out of 100 cases. A number 1 is generated by the “low” computer with a probability of 18%, in other words, in 18 out of 100 cases.

generated number	1	2	3	4	5	6	7	8
probability of the number	8%	9%	10%	12%	13%	14%	16%	18%

Table 1: The “High” Computer

generated number	1	2	3	4	5	6	7	8
probability of the number	18%	16%	14%	13%	12%	10%	9%	8%

Table 2: The “Low” Computer

As described in the beginning, it is your task to guess whether the “high” or the “low” computer is generating the numbers of the current round.

At the beginning of each round, one of the two computers is selected with equal probability. Each one of the computers has the probability 50%. The computers are selected independently over the rounds, this means that the probability that the “high” or the “low” computer is selected in a round, is always 50%. The selection of the computers is independent from which computer has been selected in the previous round.

2.2 Shown numbers as indicators of the computer

You can use the shown numbers as indicator of which computer has been selected in the respective round. For example, the number 1 is an indicator that the “low” computer has been selected in this round and is generating the numbers - however, this is not certain. As shown in table 1 and table 2, the probability that the “low”

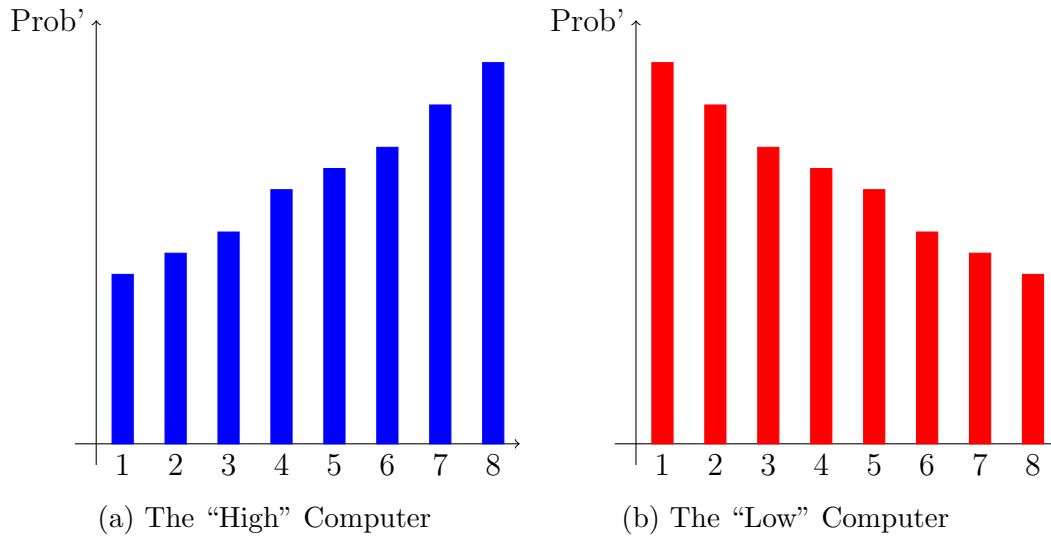


Figure 2: Graphical illustration of the probabilities with which the “high” and the “low” computer generate the numbers 1 to 8.

computer generates a number 1 is 18%, while the probability that the “high” computer generates a number 1 is only 8%.

On the contrary, when you see a number 8, it is an indicator that the “high” computer has been selected in this round and is generating the numbers. The probability that the “high” computer generates a number 8 is 18%, while the probability that the “low” computer generates a number 8 is only 8%.

In general, high numbers are an indicator that the “high” computer has been selected while low numbers are an indicator of the “low” computer having been selected. For example, the number 5 is an indicator that the “high” computer has been selected. Higher numbers, for example 6 or 7, are a stronger indicator that the “high” computer has been selected. Likewise, a number 4 is an indicator that the “low” computer has been selected, but a less strong indicator than a lower number as for example a 3 or a 2.

2.3 How we measure your guess

We will ask you for your guess which computer is generating all the numbers you see in a round. To make your guess as specific as possible, you should consider **all** numbers you see in a round; those of the first phase and those of the second phase.

Each of your guesses will have the form below:

1. Which computer is more likely?

In the first step you will be asked, which computer is generating the numbers in the current round in your opinion. This is shown in figures 3a and 3b. To answer this first question, you can click on one of the two pictured buttons and state,

which computer has been selected with a higher probability in your opinion in the respective round.

2. Your exact assessment:

In the second step, we want to know your precise assessment, i.e. how certain you feel about your guess in the first step. The tables shown in figures 3a and 3b provide some assistance. In each row you can decide between two options:

- Win 8 € if you have guessed the right computer
- Win 8 € with some probability which starts at 50% in the first row and increases by 5% per row.

One of the rows will be randomly selected for your payment. However, your choice in a row **CANNOT** influence, **which** one of the rows will be selected. Therefore, think about your choice between option 1 and option 2 very carefully in each row since every row could be selected for your payment.

An example

Assume you make the following assessment: You believe that the high computer has been selected and is generating the numbers in the respective round with a probability of 66%.

So, in the first step, for the question "Which of the two computers is more likely?" you click on the button "high".

Now, in the second step, for the question "Please specify your exact assessment", you have two options to choose from to specify your assessment:

- In the first row, you have the options "Win 8 € if "high" is right" and "Win 8 € with probability 50%". Since you believe that "high" is right with probability 66%, you should choose option 1 since this way, you win 8 € with probability 66% (instead of 50% as it would be the case with option 2).
- In the second row you have the options "Win 8 € if "high" is right" and "Win 8 € with probability 55%". Since you believe that "high" is right with probability 66%, you should choose option 1, since this way, you win 8 € with probability 66% (instead of 55% as it would be the case with option 2).
- Accordingly, you should choose option 2 in the rows where the probability of winning 8 € is 70% or higher, it is, equal or higher to the probability with which you believe that the computer you think has been chosen is right.

Time left to submit your guess: 0:27

Round 1 - First Guess

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

Option 1		Option 2
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 50 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 55 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 60 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 65 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 70 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 75 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 80 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 85 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 90 %
Win 8€, if "high" is true	<input type="radio"/>	Win 8€ with probability 95 %

Next

Time left to submit your guess: 0:23

Round 1 - First Guess

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

Option 1		Option 2
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 50 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 55 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 60 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 65 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 70 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 75 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 80 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 85 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 90 %
Win 8€, if "low" is true	<input type="radio"/>	Win 8€ with probability 95 %

Next

(a) The monitor after you clicked "High"

(b) The monitor after you clicked "Low"

Figure 3: The assessment monitor of the first guess in the beginning of a round, after you have clicked "high" or "low"

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

	Option 1		Option 2	
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>		Win 8€ with probability 50 %
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>		Win 8€ with probability 55 %
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>		Win 8€ with probability 60 %
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>		Win 8€ with probability 65 %
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>		Win 8€ with probability 70 %
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>		Win 8€ with probability 75 %
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>		Win 8€ with probability 80 %
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>		Win 8€ with probability 85 %
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>		Win 8€ with probability 90 %
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>		Win 8€ with probability 95 %

Next

Figure 4: The guessing screen from the example with a guessed probability of 66% for the “high” computer. The button for “High” and the button of option 2 with winning probability 70% has been clicked.

So, as soon as the probability of winning in option 2 is higher than your certainty of your guess (whether the high/low computer has been selected), you should choose option 2. This is illustrated in figure 4.

Please notice the fill-in assistance: The fill-in assistance will automatically choose option 2 in all the following choices under option 2 with a higher winning probability than the one you have chosen (it is, all the rows under the row where you have chosen option 2 for the first time), since the winning probabilities are increasing by 5% per row.

After you have chosen option 2 with a winning probability of 70%, all following rows with a higher probability than 70% in option 2 will be automatically chosen for you.

On the other hand, when you think that it is more likely that the numbers in the re-

spective round are generated by the “low” computer, you click on the button “low” in the first step. For the second step, you proceed as described above and compare for each row, whether you prefer option 1 or option 2. You can indicate your exact assessment as described above. The only difference lies in option 1, as illustrated in figure 3b: You win 8 € if the “low” computer has been selected.

Reminder:

- In the first step you indicate which computer you think is more likely
- In the second step, you make a more exact assessment:
 - Therefore, you should read the table row by row and compare option 1 to option 2 in each row to decide which option you prefer in the respective row.
 - This is very important, since every one of your decisions is relevant for your payment and determines, how much you will earn in this experiment. Therefore, please think about your choices very carefully.
- As soon as the winning probability in the second step under option 2 is higher than your certainty of your guess (whether the high or low computer has been selected), you should choose option 2
- The fill-in assistance will automatically choose option 2 for you in all the following choices with a higher winning probability in option 2 than the one where you have chosen option 2 for the first time.

3 The sequence of each round

In the following we explain the procedure of the experiment to you by guiding you through the sequence of a round. In this part of the experiments, 12 rounds will be played. Each round consists of a number of guesses and phases. In the phases of a round, you see numbers which you can use as indication for which computer has been selected in the respective round. The sequence of a round is illustrated in figure 5.

3.1 A computer is randomly selected

At the beginning of each round, one of the two computers (it is, the “high” or the “low” computer) is randomly selected. Each of the computers (high or low) has the same chance to be selected. Thus, the probability for the “high” or the “low” computer is 50% in each case at the beginning of a round. You will see a screen which points out that a new round has started and once again one of the two computers (“high” or “low”) has been randomly selected.

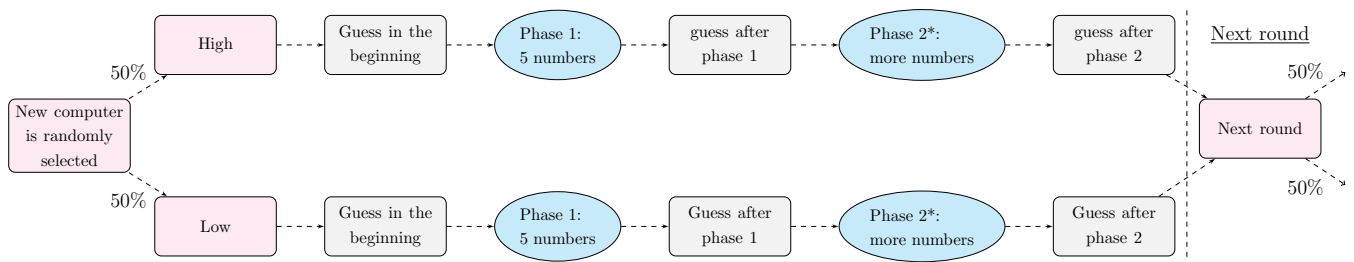


Figure 5: Sequence of a round
 (* Phase 2 can occur in 2 versions)

3.2 Guess at the beginning of a round

In each round, at the beginning of the round, the “high” or the “low” computer will be randomly selected with a probability of 50% each. This happens randomly at the beginning of each round.

At the beginning of a round, before you see any numbers, we will ask you for your guess which computer has been randomly selected. We do this to make sure that you know you are at the **beginning** of a round. You have 30 seconds to make your guess.

Reminder: If you do not feel confident how to fill out the assessment screen or do not know when to choose option 1 or option 2 in a row, please read section 2.3 “How we measure your guess” again.

3.3 Phase 1

In phase 1 you will see 5 numbers, as illustrated in figure 6. Those numbers are generated by the computer which has been randomly selected at the beginning of the current round; for example “5 7 3 2 2” or “7 7 6 4 2”. You have 30 seconds time to look at the numbers and to form your assessment. After 30 seconds, the numbers will disappear and you will be directed to the next screen. On the next screen, you will be asked to indicate your guess as described above.



Figure 6: Screenshot of Phase 1

3.4 Guess after phase 1

After phase 1, we will ask you again to make a guess which computer has been selected at the beginning of the round and is now generating the numbers. You can use the numbers from phase 1 as indication of the randomly selected computer. Again, you will indicate your guess in the table from figures 3a and 3b, at this, you will see your assessment from the first guess as default setting. However, you can change this assessment as you like. You have 30 seconds time to make indicate your guess and to make it more precise.

Reminder: If you do not feel confident how to fill out the assessment screen or do not know when to choose option 1 or option 2 in a row, please read section 2.3 “How we measure your guess” again.

3.5 Phase 2

In phase 2, you will see up to 7 additional numbers. These numbers are generated by the computer which has been randomly selected at the beginning of the current round. There are two versions of phase 2 which can switch randomly from round to round.

Phase 2, Version 1

In version 1 of phase 2, you can reveal up to 7 additional numbers. Again, those numbers are generated by the computer which has been randomly selected at the beginning of the current round and has already generated the 5 numbers from phase 1 of the current round. You can only see one number at a time: When you uncover the next number, the number shown until then will disappear. You have no possibility to go back to this number.

The first number appears as soon as phase 2 starts. When you want to see the next number in this version of phase, you can click “Next”. You will be redirected to a screen as in figure 7.

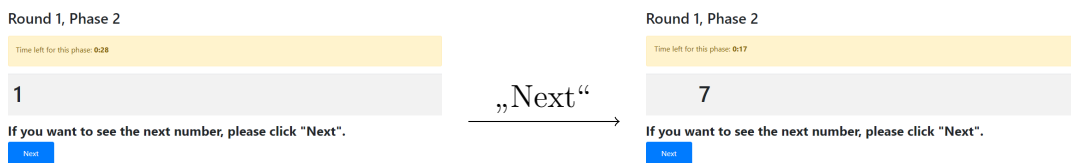


Figure 7: Screen of phase 2, version 1

Please notice: As soon as you click “next”, the currently displayed number will disappear. You have **no** possibility to go back to the previous screen to see this number again.

After 30 seconds in phase 2 and no matter whether you have seen all 7 numbers, you will be redirected to the screen for the guess after phase 2. You will have 30 seconds in phase 2 in total and cannot proceed earlier. Thus, consider carefully how you want to allocate your time between the 7 numbers that you can uncover in total.

Phase 2, Version 2

In version 2 of phase 2 you will be shown up to 7 additional numbers. Again, those numbers are generated by the computer which has been randomly selected at the beginning of the current round and has already generated the 5 numbers from phase 1 of the current round.

The additional numbers appear one after another on your monitor. In this version of phase 2, you **cannot** control the display of the next numbers. Instead, the numbers will be shown automatically. Differently to version 1, the shown numbers will not disappear again: The previous numbers will be still visible. An example is shown in figure 8.

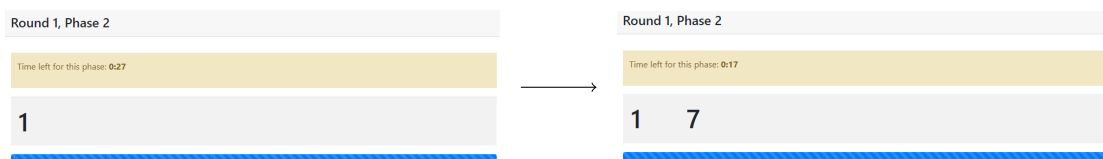


Figure 8: Screen of phase 2, version 2

After 30 seconds have passed, you will be redirected to the next screen to make your guess after phase 2. Note that you have 30 seconds time but it can happen that you see less than 7 numbers in these 30 seconds.

3.6 Guess after phase 2

After phase 2, we will ask you again to make a guess which computer has been selected at the beginning of the round. You can use the numbers from phase 1 and phase 2 as indication of the randomly selected computer. Again, you will indicate your guess in the table from figures 3a and 3b, at this, you will see your assessment from the first guess as default setting. However, you can change this assessment as you like. You have 30 seconds time to indicate your guess and to make it more precise.

Reminder: If you do not feel confident how to fill out the assessment screen or do not know when to choose option 1 or option 2 in a row, please read section 2.3 “How we measure your guess” again.

3.7 Next round

After your guess after phase 2, a new round will start and a new computer (the “high” or the “low” one) will be randomly selected and will be generating the numbers in the new round.

4 How you will get paid

For this part of the experiment, you play 12 rounds with 3 guesses each per round. From these guesses, we will randomly select 3 of your guesses:

One guess at the beginning of a round, one guess after you have seen 5 numbers in phase 1, and one guess at the end of a round after you have seen up to 7 numbers in phase 2. Each of these randomly selected guesses will come from a different round. Subsequently, from each of these guesses, a row will be randomly selected in the corresponding decision table. Your choice in this row will determine your payment:

1. if you chose option 1, you will win 8 € if you guessed correctly whether it was a “high” or “low” computer generating the numbers of the round;
2. if you chose option 2, you will win 8 € with the probability specified in the row we randomly selected.

5 Control Questions

1. In the first guess of a round (before you have seen the numbers of phase 1), what is the probability that the “high” computer has been selected?

Answer: The probability is _____ percent.

2. In the first guess of a round (before you have seen the numbers of phase 1), what is the probability that the “low” computer has been selected?

Answer: The probability is _____ percent.

3. Suppose that in the previous round, you have seen the numbers

1, 2, 2, 3, 3, 1, 4, 7.

Now, in the first guess of the next round (before you have seen the numbers of phase 1), what is the probability that the “high” computer was selected for this round? Why?

Answer: The probability is _____ percent.

Please explain: _____

4. What do you choose in the table when you believe that the “low” computer is right with a probability of 72%? Please draw your choice in the table below.

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

Option 1		Option 2
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 50 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 55 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 60 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 65 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 70 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 75 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 80 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 85 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 90 %
Win 8€, if "low" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 95 %

Next

5. Assume that you think at the beginning of a round, that the probability for the “high” computer is 50%. Please draw in the table below, how the screen should look like before you would click “next”.

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

Option 1		Option 2
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 50 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 55 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 60 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 65 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 70 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 75 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 80 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 85 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 90 %
Win 8€, if "____" is true	<input type="radio"/> <input type="radio"/>	Win 8€ with probability 95 %

Next

6. Take a look at the following example: After you have seen the numbers, you believe that the “high” computer has been selected with a probability of 85%. What has not been filled in correctly in the following screen?

Which Computer is more likely that has been selected?

"high" "low"

Please specify your assessment:

	Option 1		Option 2	
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>	Win 8€ with probability 50 %	
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>	Win 8€ with probability 55 %	
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>	Win 8€ with probability 60 %	
Win 8€, if "high" is true	<input checked="" type="radio"/>	<input type="radio"/>	Win 8€ with probability 65 %	
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>	Win 8€ with probability 70 %	
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>	Win 8€ with probability 75 %	
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>	Win 8€ with probability 80 %	
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>	Win 8€ with probability 85 %	
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>	Win 8€ with probability 90 %	
Win 8€, if "high" is true	<input type="radio"/>	<input checked="" type="radio"/>	Win 8€ with probability 95 %	

Next

Answer: _____

Chapter 3

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

JOINT WITH LORENZ GÖTTE AND ZHI HAO LIM

3.1 Introduction

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

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

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

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

³This view is also consistent with the Köszegi and Rabin (2006) model of reference-dependent preferences, in which reference points are given by recent expectations: Goals may affect expectations, and thus create a reference point.

around them.^{4 5}

Goals may change over time. In management, business objectives may change due to economic shifts or evolution of a firm (e.g. Fisher et al., 2016; Kennerley and Neely, 2003). If goals directly affect individuals’ motivation, changing them poses additional challenges. For instance, consider two goals, M and H , where M is a moderate goal which is easier to reach than H which is a hard goal. Consider two individuals, one of which starts with the moderate goal M and the other with the hard goal H . If goals serve as reference points, it may well be that individuals exert more effort to reach M , while diminishing sensitivity leads them to exert less effort to try and reach H (Heath et al., 1999). Consider now moving both individuals to an intermediate goal I , which is halfway between M and H . For individuals starting out at M , the goal has become harder. However, because their new reference point is (near) M , loss aversion incentivizes them to work hard in order to try and reach I . By contrast, individuals who started out with H as the goal now feel that the new goal I is easier relative to their reference point. Working towards it feels like a gain, and consequently, effort will be lower. Thus, even though all individuals now face the same goal I , those who started out with the moderate goal M perform better in the long run than those who started out with the hard goal H .

In this paper, we conduct a field experiment to test this central prediction. The experiment is set in the context of household water conservation by individuals. Our sample consists of over 600 students in two residential colleges at the National University of Singapore. It is important to note that residents are not charged for any of their water (or energy) use. Thus, the setting allows us to focus on the effects of goals on non-pecuniary motivations.

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

In phase 1 of the experiment, we assigned subjects to one of four conditions. In our two key experimental conditions, subjects are encouraged to keep their water use per shower below a specified target: One condition received an 18L target, the other condition received a 28L target. Baseline use was around 32L on average, thus making the former a hard goal to reach (requiring halving of the water use to reach it), and the latter a moderate goal (requiring to save only about 15 percent of water per shower). Smart shower heads provided feedback relative to the goal: They shine in a green light at the beginning of a shower, and subsequently change color to yellow, followed by orange and then red with increasing water use. Finally, the shower heads display a blinking red light when the water volume exceeds the specified shower goal. These thresholds were clearly communicated through posters in each shower stall. To control for potential moral suasion, a third experimental condition only received a poster encouraging them to keep water use below 28L,

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

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

but no feedback to track their behavior. Finally, a fourth experimental condition received neither feedback nor moral suasion.

In phase 2 of the experiment, we change both goal conditions as well as the moral suasion condition to an intermediate goal of 24L per shower.

Overall, we find that moral suasion alone did not induce any significant effect on water use. However, assigning a goal combined with real-time feedback led to large and significant conservation effects of around 14 – 19 % of baseline shower water use, compared to the control group and the moral suasion group. Interestingly, we do not find differences in the average treatment effects between the *18L GOAL* and the *28L GOAL* groups in phase 1. However, when both groups are assigned to the intermediate (24L) goal in phase 2, we observe a stark divergence in average treatment effects driven by the *18L GOAL* group: Subjects assigned the hard (18L) goal systematically underperform under the 24L goal relative to their counterparts who were initially assigned the moderate (28L) goal. In addition, we find evidence of heterogeneity: During phase 1, high baseline users in the *28L GOAL* group show larger conservation effects than their counterparts in the *18L GOAL* group, where conservation effects are uniform across baseline use. Interestingly, this heterogeneity persists even in phase 2, even though both groups now face the same 24L goal and feedback.

Thus, in phase 1 of our experiment, we find, as e.g. Goerg et al. (2019) or Agarwal et al. (2018) did, that harder goals do not necessarily lead to better outcomes. As Agarwal et al. (2018), we further find heterogeneity in the interaction of reaction towards the real-time feedback (as shown in average treatment effects) with the baseline use between the 18L group and the 28L group. However, we additionally show that these heterogeneities carry over to the next phase when both goals are changed to the same level. We are, to our knowledge, the first to show that setting an overly ambitious goal is associated with lasting (potentially detrimental) effects on effort and performance, even after the goal had been adjusted to an intermediate level. Taken together, our findings suggest that initial goal assignment is critical and leads to long-term effects on performance even after the goal had been changed. In our setting, this effect was particularly pronounced for high-baseline users: The hard goal muted their conservation effects right from the outset, and this effect also persisted in phase 2 under the new intermediate goal.

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

But none of these measures specifically addresses the potential pitfalls that might

arise when goals change over time, except for the fact that in business environments which might be changing, performance goals might actually prevent learning. Locke and Latham (2002) propose the use of learning goals instead of performance goals in “complex environments”, but this may not always be feasible outside business scenarios.

The economics literature is subdivided into literature on exogenous and endogenous goals. Endogenous goals are self-chosen goals (see e.g. Brookins et al. (2017), who provide a good overview of the existing literature on endogenous goals). Harding and Hsiaw (2014) find that for self-set goals on resource conservation, realistic goals will lead to more savings than very low or unrealistic high goals.⁶ Our paper falls under the category of exogenous goals, which are set by another party. In general, agents respond to exogenously set goals and exogenous goals have a positive effect on performance, in particular when they are attainable (Corgnet et al., 2015; Gómez-Miñambres, 2012; Wu et al., 2008). However, to our knowledge, the economics literature on the goal-performance relationship for exogenous goals is mostly theoretical or empirically studied in the lab. We add to these findings with a setting that allows us to look at choices without monetary incentives. Moreover, in the economics literature, goals can be seen as reference points as mentioned above (Heath et al., 1999; Kahneman and Tversky, 1979; Tversky and Kahneman, 1991). If goals can have the properties of reference points, individuals might experience related effects such as loss aversion and diminishing sensitivity around them which can cause challenges when goals are changed. And as mentioned above, the literature on the effect of changing goals (and the subsequent effects of hard vs. moderate goals) is sparse. Our contribution to the literature on goal-setting is twofold: First, by examining the goal-performance relationship for exogenous goals in a randomized field experiment. Second, by looking at “dynamic goals”, i.e. how the goal-performance relationship plays out once goals are changed. In particular, we test whether individuals who start with a moderate goal will perform better once goals are changed to the same intermediate goal compared to individuals who start with a hard goal.

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

One promising intervention is to supplement goal setting with the provision of information feedback; this has been shown to significantly reduce resource consumption (e.g. Abrahamse et al., 2005; Attari et al., 2010; Becker, 1978; Harding and Hsiaw, 2014; Tiefenbeck et al., 2018). We contribute to this strand of literature by carefully examining the complementarities of goals and real-time feedback in a randomized controlled trial using smart shower heads displaying lights for feedback as tools for resource conservation, with a special emphasis on the effects of initial goal difficulty. Our setting further allows us to test these behavioral interventions in the absence of monetary incentives.

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

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

3.2 Experimental Design

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

3.2.1 Background

Each residential college consists of 21 stories with over 600 rooms in total, providing accommodation to local undergraduates, international exchange students, and a small group of faculty members. At the beginning of each semester, students can opt for their preferred room type (i.e. single corridor room or single room in shared apartment) on either mixed or single-gender floors. Our pool of subjects comprises mainly incoming freshmen and excludes all faculty members. Figure 3.1 displays some photos of the experimental site at Cinnamon and Tembusu colleges.

In total, 324 HYDRAO smart shower heads were installed in all designated bathrooms at Cinnamon and Tembusu colleges. Note that there are two types of bathrooms on each floor: apartment and common bathrooms (see Figure 3.8). Residents who live in a shared apartment have access to their own apartment bathroom, while those who stay in the single corridor rooms use the common bathrooms.⁷ From anecdotal evidence, residents typically store their toiletries in one particular bathroom, and hence it is safe to assume that the majority use the same bathroom for showers. In light of this, we chose to randomize at the residence \times floor \times bathroom type level. Each unit of randomization consists of between 4 and 6 shower heads that receive the same treatment assignment, shared by 18 residents on average.

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

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



(a) *Tembusu and Cinnamon colleges*



(b) *Single room*



(c) *Bathroom with 2 shower facilities*

Figure 3.1: *Experimental site*

3.2.2 HYDRAO smart shower head

The smart shower head is engineered by HYDRAO, a French water-technology and data startup. During a shower event, the smart shower head displays a colored light, which changes in real-time based on the level of water use. This provides users with real-time feedback about their shower water use. The exact thresholds and corresponding colors can be configured, which allows us to implement different (exogenous) goals for our treatment groups.

For each shower event, the smart shower head can collect real-time data when it is connected to the server via WiFi. As a safeguard, there is an internal memory of 200 shower events. This means that if a particular shower fails to be transmitted real-time, the data will be stored and transmitted as an offline shower event as soon as the WiFi connection is re-established. If the shower is interrupted for a short duration of time (e.g. for soaping purposes), the smart shower head will still consider it the same shower event as long as the interruption is under 2 minutes. Beyond 2 minutes, the shower head will assume that a new shower event has started.

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

3.2.3 Treatment assignment

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

In the baseline period, no intervention was implemented in all shower facilities. This means that there were no poster and no color-display feedback through smart shower heads. The primary objective was to collect information on pre-experimental showering behavior of the residents. Our baseline data contain observable characteristics from each shower head, such as water use per shower, number of showers per day and flow rate. We use this information to conduct randomization checks in the following section.

For the intervention, we implemented four experimental groups: One control and three treatments. These assignments were permanent throughout the rest of the experiment. The *Control* group received neither the shower poster nor real-time feedback in both phases. The *Moral Suasion (MS)* group received a shower poster appealing to users to keep their water use under a specified level, but no feedback through the shower heads. The shower poster references a goal of 28L in phase 1, and subsequently 24L in phase 2. The *18L GOAL* group received a shower poster and real-time feedback that corresponds to the goal of 18L in phase 1, and thereafter 24L in phase 2. The *28L GOAL* group received a shower poster and real-time feedback that corresponds to the goal of 28L, which was similarly switched to 24L in phase 2. The treatment groups thus initially received different (exogenous) goals in phase 1 but identical goal in phase 2, which were conveyed through the shower posters. Table 3.1 summarizes the key features of the experimental groups and Figures 3.9, 3.10 and 3.11 present the respective shower posters.

Table 3.1: Summary of treatment assignments

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

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

Further, for the *18L GOAL* and *28L GOAL* groups, we programmed the smart shower heads to display real-time feedback, in the form of colored lights (resembling the traffic light system) that correspond to a set of thresholds. The shower head would display a green light at the start of each shower, then progress to yellow, orange and red with increasing water consumption. When the volume of water use exceeds the goal, the shower head would begin to display blinking red light. Figure 3.3 depicts a shower head assigned to the *18L GOAL* group in action. In the first panel, the shower head is displaying a green light, indicating that water use up to that particular point is below 12 liters. For ease of interpretation, we also show the composition of feedback lights for a typical shower from start to end, based on the assigned experimental group, in Figure 3.2.





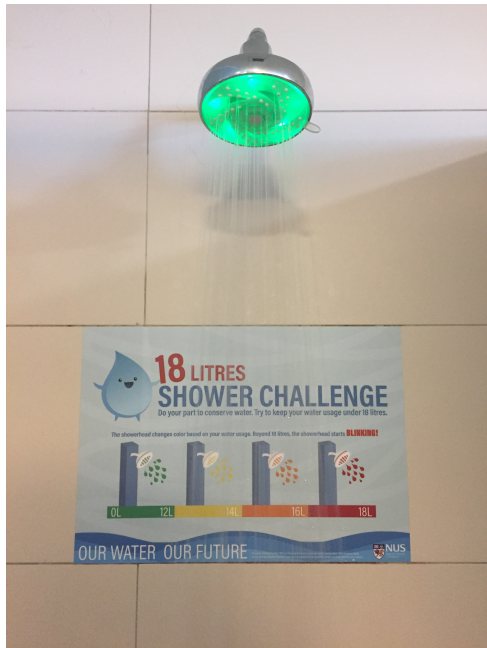
	Phase 1	Phase 2
Control	No poster, no feedback	No poster, no feedback
Moral Suasion	28L goal on poster (Figure 4.9(a))	24L goal on poster (Figure 4.9(b))
18L GOAL	18L goal on poster (Figure 4.10(a)) + feedback up to 12 14 16 18 20 22 24 26 28 more than 28 	24L goal on poster (Figure 4.10(b)) + feedback up to 12 14 16 18 20 22 24 26 28 more than 28 
28L GOAL	28L goal on poster (Figure 4.11(a)) + feedback up to 12 14 16 18 20 22 24 26 28 more than 28 	24 goal on poster (Figure 4.11(b)) + feedback up to 12 14 16 18 20 22 24 26 28 more than 28 

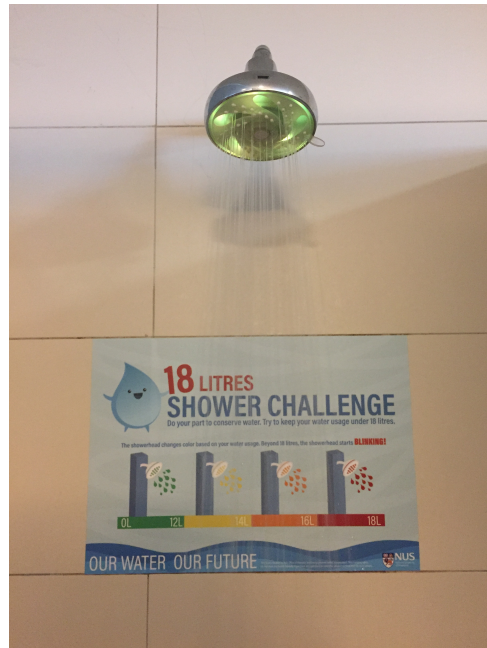
Figure 3.2: Experimental groups in phase 1 vs phase 2.

Our experimental design allows us to identify the effect of moral suasion alone

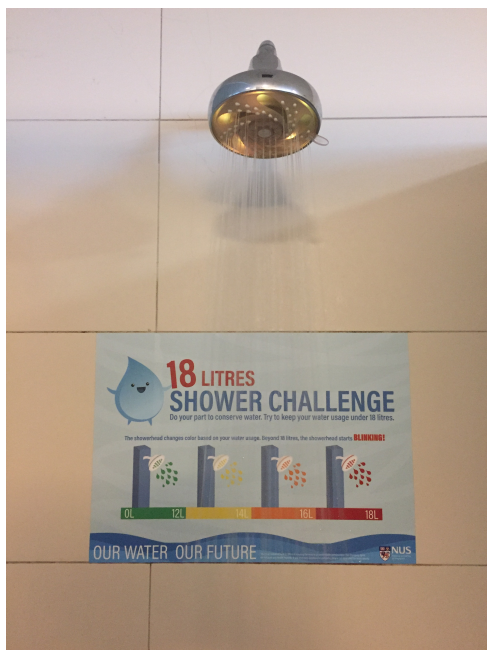
and the marginal effect of real-time feedback (on top of moral suasion) under different (exogenous) goals. In addition, the implementation of different goals in phase 1 but the same goal in phase 2 allows us to study the relationship between goal difficulty and effort provision in a dynamic setting.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



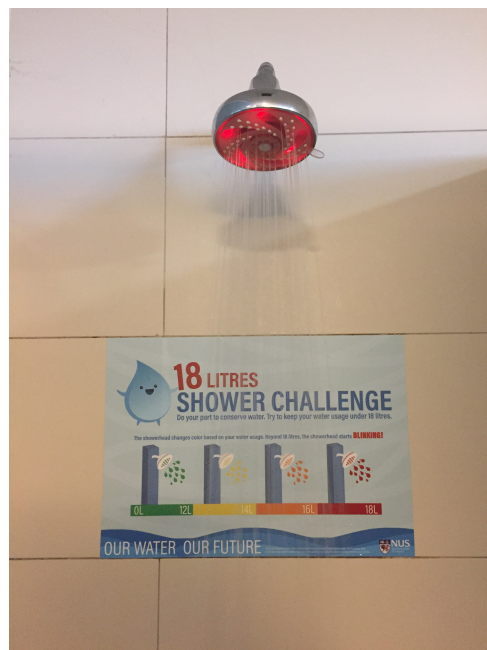
(a) 0L to 12L



(b) 12L to 14L



(c) 14L to 16L



(d) 16L to 18L

Notes. Beyond 18L, the shower head starts to display blinking red light.

Figure 3.3: Implementation of 18L GOAL group

3.2.4 Behavioral predictions

Our experimental setup allows us to test three predictions derived from the literature on goal setting and reference-dependent preferences (Heath et al., 1999; Köszegi and Rabin, 2006).

If goals act as reference points, then the psychological utility of reaching the goal depends on loss aversion and diminishing sensitivity. An ambitious goal may elicit strong effort, because it puts an individual firmly in the loss domain and mitigating the loss has a high marginal value. However, there is a counteracting force: If goals are too ambitious, they may elicit lower effort from individuals because of diminishing sensitivity relative to the reference point (Heath et al., 1999).

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

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

Several different mechanisms of how goals affect reference points could lead to hypothesis 1: It could be that goals directly serve as reference points and thus create the above pattern, as is argued in Heath et al. (1999). However, it could also be that reference points are driven by recent expectations, as in Köszegi and Rabin (2006), and goals merely serve to influence those expectations. The main feature we are interested in testing is how individuals respond when the difficulty of goals is changed over time. For the 28L condition, phase 2 of the experiment introduces a tougher goal of 24L. How the two groups respond to introducing the new goal depends on whether and how goals affect reference points.

If goals directly act as reference outcomes, then the introduction of a common 24L goal in phase 2 for the groups who formerly had the 18L and 28L goal, respectively, should lead to the same reference point, and hence to the same outcomes. We summarize this as

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

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

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

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

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

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

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

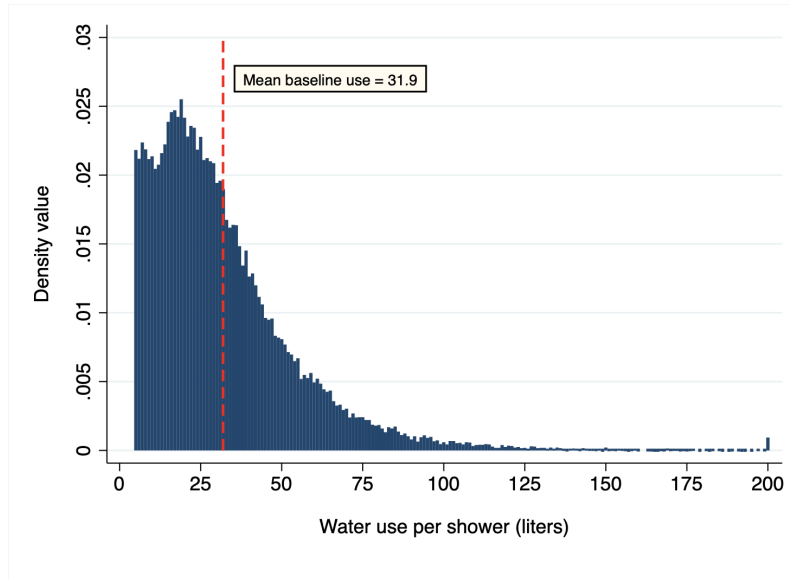
The mechanism derived from changes in reference outcomes leads to similar predictions as when there is habit formation in the sense of Stigler and Becker (1977): In their model, such an effect could result from habit stock building up over time, and thus gradually reducing the marginal utility of showering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

3.3 Analysis and Interpretation

Our data comes from the randomized field experiment described above. In total, we recorded 128,323 shower instances (of strictly more than 4 liters), from 301 working shower heads. The observations can be classified into two types: live showers and offline showers. The former refer to shower instances where data was transmitted real-time (at the point of showering) from the shower heads to our server — this gives us information about the actual date and time each shower event took place. In contrast, the latter refer to shower instances that were transmitted with a time lag, and so we are unable to accurately pinpoint the occurrence of these shower events. To increase precision of our estimates, we thus consider the sample of 116,891 live showers (instead of all recorded showers) as our primary data source for analysis.

In the baseline period, we record a stable pattern of around 1,300 live showers on a regular weekday, and about half the number on weekends. On the intensive margin, we observe a right-skewed distribution of baseline water use per shower, with a mean of 31.9 liters (see Figure 3.4).

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



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

Figure 3.4: *Distribution of baseline water use per shower*

3.3.1 Randomization checks

To begin, we perform balance tests to support the integrity of the randomization. Table 3.2 presents a comparison across different treatment groups, relative to the control, on the baseline averages of key observables. It is apparent that balance of treatment is attained as almost all observables, in particular water use per shower and fraction of live showers, do not differ across groups. There is only slight statistical difference in the number of days since last transmission between the control and *28L GOAL* group.¹⁰ This is largely driven by a single shower head in the control group which rarely records live shower events, and thus does not constitute a cause for concern.¹¹ We conclude that our experimental groups are well-balanced and interpret any observed differences during intervention as causal treatment effects.

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

¹¹In particular, the shower head last transmitted shower data 24 days before the start of phase 1, possibly due to poor Wi-Fi coverage in the bathroom. Out of the 238 observations recorded by the shower head during the entire experiment, only 11% are classified as live showers.

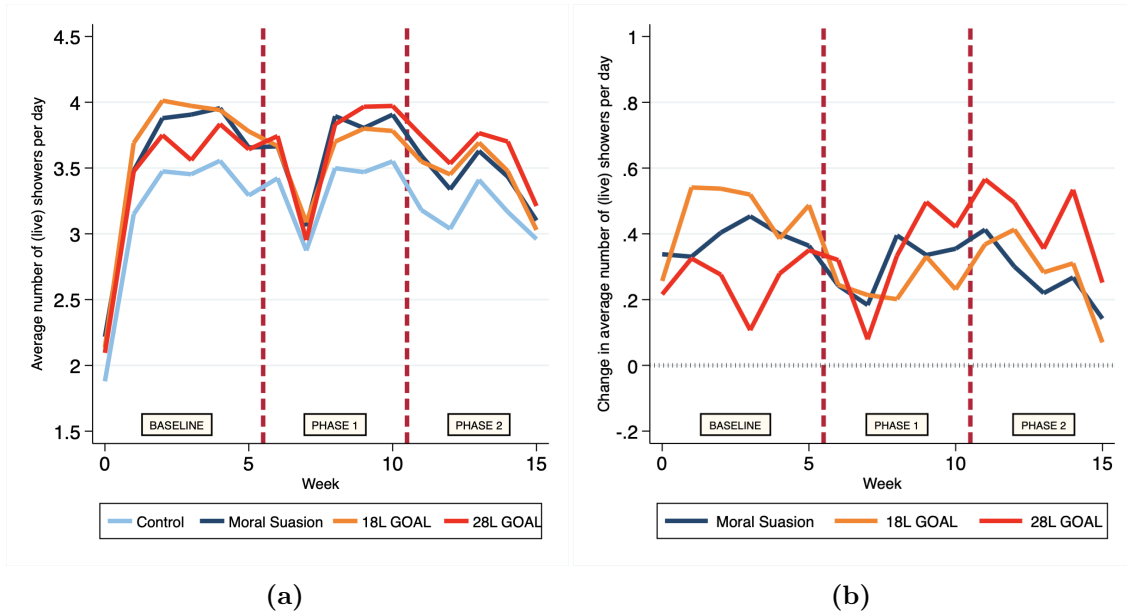
Table 3.2: *Randomization checks*

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

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

3.3.2 Descriptive evidence

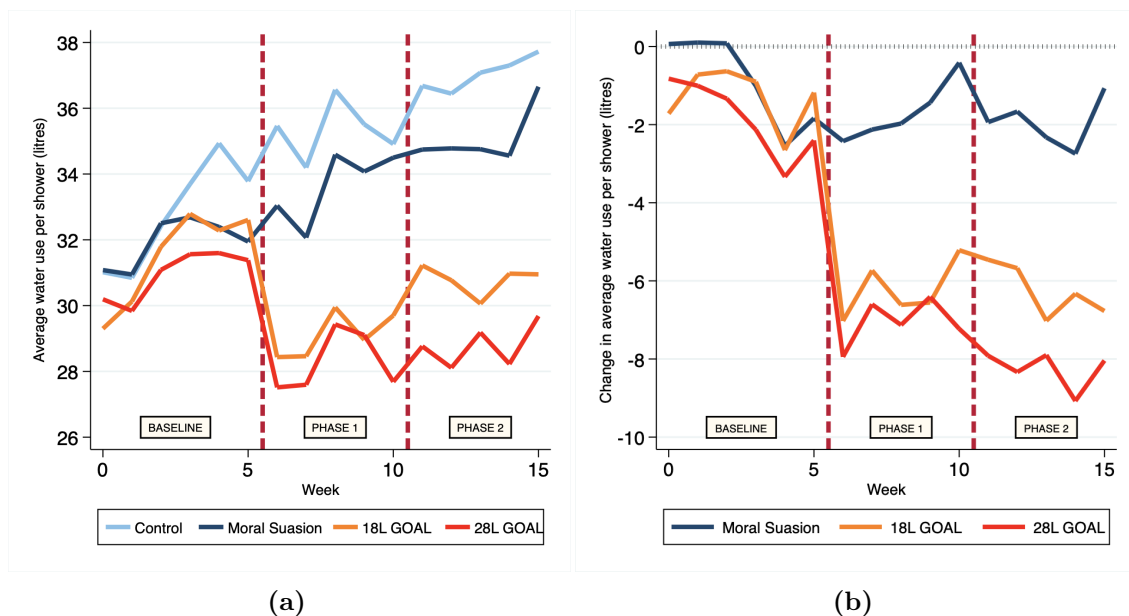
In the context of goal-setting, our empirical analysis compares how moral suasion and real-time information feedback influence shower behavior on the extensive and intensive margins. We break the full sample into three distinct time periods: baseline, phase 1 and phase 2. Recall that in phase 1, the *18L GOAL* and *28L GOAL* groups received moral suasion and real-time feedback which referenced different goals (i.e. 18L vs. 28L), and subsequently in phase 2, both groups were moved to the common goal of 24L. Figures 3.5 and 3.6 provide descriptive evidence of how our treatments impacted daily number of showers (extensive margin) and water use per shower (intensive margin) over time, respectively.



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

Figure 3.5: *Extensive margin of shower behavior by experimental groups*

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



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

Figure 3.6: *Intensive margin of shower behavior by experimental groups*

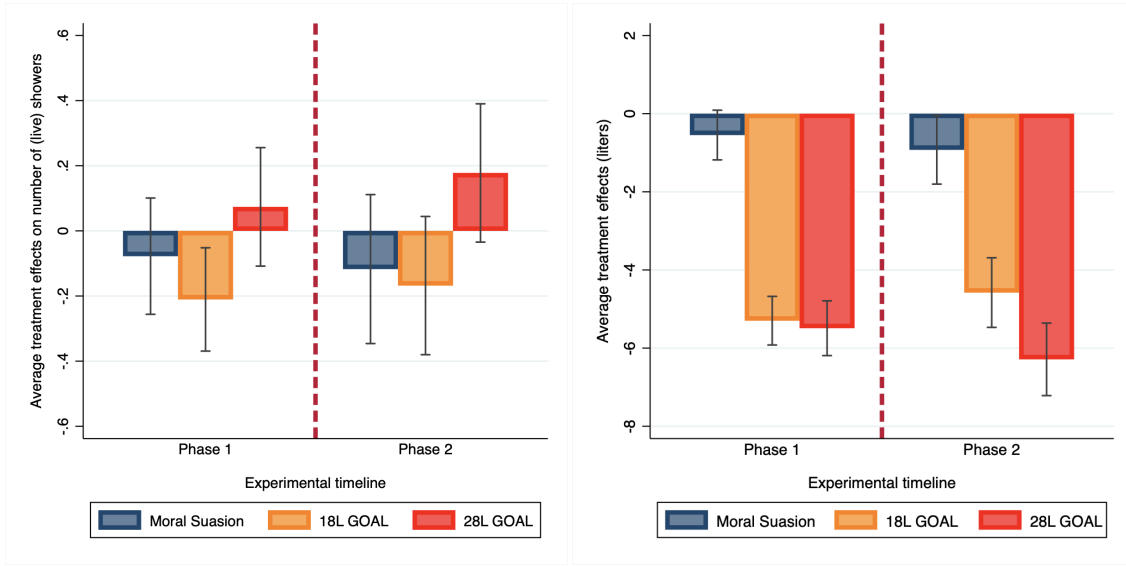
On the intensive margin, we observe that all experimental groups have near-similar levels of mean water use per shower during the baseline period, consistent with our preceding randomization checks. For the control group, there is a visible upward trend of mean water use per shower over time; the *MS*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

To augment our analysis, Figure 3.7 displays the difference-in-differences estimates of the average treatment effects for the *MS*, *18L GOAL* and *28L GOAL* groups. We examine whether our treatments have an impact on the extensive and intensive margins of shower behavior in the left and right panels, respectively. The estimates are obtained by taking the difference between the outcome of interest in phase 1 and the baseline period, and similarly the difference between phase 2 and the baseline period.

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

responses to our respective treatments.

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



(a) *Extensive margin*

(b) *Intensive margin*

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

Figure 3.7: *Average treatment effects by experimental groups*

3.3.3 Estimation strategy

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

$$y_{ith} = \alpha_i + \lambda_t + \gamma_h + \left(\beta_{MS1}MS_i + \beta_{18L1}18L_i + \beta_{28L1}28L_i \right) \times \text{PHASE1}_{ith} + \left(\beta_{MS2}MS_i + \beta_{18L2}18L_i + \beta_{28L2}28L_i \right) \times \text{PHASE2}_{ith} + \epsilon_{ith} \quad (3.1)$$

where y_{ith} is the outcome variable of interest, e.g. water use per shower for device i on day t and hour h . α_i is the device fixed effect, λ_t is the day fixed effect and γ_h is the hour-of-day fixed effect. MS_i is a dummy variable that equals one for the *MS*, *18L GOAL* and *28L GOAL* groups that all receive moral suasion, in the form of a shower poster. The $18L_i$ and $28L_i$ variables are indicators for being assigned to the *18L GOAL* and *28L GOAL* groups respectively. Note that the MS_i variable is defined to be one for all three treatment groups, instead of only the *MS* group. This is because the *18L GOAL* and *28L GOAL* groups both receive moral suasion, and therefore the *MS* group serves as the relevant comparison for identifying the marginal effects of real-time feedback (on top of moral suasion). $PHASE1_{ith}$ is a dummy variable that equals one for the period when the initial shower goals (i.e. either 18L or 28L) were introduced, whereas $PHASE2_{ith}$ is a dummy variable that equals one for the latter period when the shower goal is changed to 24L.¹² $\epsilon_{i,t,h}$ is the random error term and standard errors are clustered at the residence \times floor \times bathroom type level (i.e. unit of randomization).

Our preferred specification includes device fixed effects to account for time-invariant differences across residents who are assigned to different experimental groups, as well as day and hour-of-day fixed effects to control for aggregate patterns in weather and lifestyle over the course of the experiment. The coefficients of interest are the respective β terms, which represent difference-in-differences estimates for each of the treatment conditions relative to the control. The average treatment effects are identified from within-device variation over time, controlling for aggregate hourly and daily shocks. To elaborate, β_{MS1} can be interpreted as the average treatment effect of moral suasion on water use per shower in phase 1 relative to the baseline, and β_{MS2} gives the corresponding average treatment effect in phase 2 relative to the baseline. Similarly, the coefficients on the interacted $18L_i$ and $28L_i$ variables represent the effect of real-time feedback over and above the effect of moral suasion in each respective phase. Table 3.3 presents the results.

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

Next, when the shower goal is adjusted to 24L in phase 2, we observe a divergence of the average treatment effects. As shown in column 2, while the 18L condition fared worse, the 28L condition responded more strongly to real-time feedback in phase 2

¹²To be specific, for the *MS* group, $PHASE1_{ith}$ equals one from 4PM, September 16 to 5PM, October 22, whereas $PHASE2_{ith}$ equals one from 5PM, October 22 onwards. For the *18L GOAL* and *28L GOAL* groups, we use the exact date and time the feedback was in place to define $PHASE1_{ith}$ and $PHASE2_{ith}$ variables respectively.

relative to phase 1. In particular, we are able to marginally reject the null hypothesis of equality between the 18L and 28L conditions in phase 2 ($p = 0.061$). Comparing between the 18L and 28L conditions, we further test for equality of the change in treatment effects from phase 1 to 2, and can easily reject the null hypothesis at the 1% level ($p = 0.008$). Thus, our data soundly rejects Hypothesis 2a that the shower goals serve directly as reference points, as this interpretation would have yielded the same outcomes for both the *18L GOAL* and *28L GOAL* groups. On the flip side, our results lend credence to Hypothesis 2b that reference points are affected by recent expectations or lagged outcomes, as observed from the underperformance in conservation efforts by the *18L GOAL* relative to the *28L GOAL* group.

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

Table 3.3: *Effects of moral suasion and real-time feedback on water use per shower*

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

Notes. This table shows the effects of moral suasion and real-time feedback (based off different shower goals in each phase) on water use per shower. The results are obtained by estimating equation (3.1) that includes controls for device, day and hour dummies. For ease of interpretation, we split the estimates from a single regression into 2 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. *live* showers) is reported in the last row. Standard errors clustered at the residence \times floor \times bathroom type level in parentheses.

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

Next, we examine how subjects adjusted their showering behavior on the extensive margin. Given that real-time feedback, coupled with moral suasion, induced large conservation effects of roughly 14% – 19%¹³, it is pertinent to consider if the savings were offset by subjects taking more showers each day. On the flip side, if the treatments had induced subjects to take fewer showers each day, this might generate negative externalities in the form of hygiene problems. In addition, there may be attrition bias, where subjects drop out of the study non-randomly. To test this formally, we re-estimate equation (3.1) without hour-of-day fixed effects, this time

¹³This corresponds to between 4.5 and 6.2 liters per shower, as is evident from Table 3.3. The lower bound of 4.5 liters per shower is given by the sum of estimates $\hat{\beta}_{MS2}$ and $\hat{\beta}_{18L2}$, while the upper bound of 6.2 liters is given by the sum of estimates $\hat{\beta}_{MS2}$ and $\hat{\beta}_{28L2}$.

using number of live showers per shower head per day as the outcome variable.

Table 3.4 presents the results, with all point estimates being statistically insignificant. This is consistent with the descriptive evidence presented above (see Figure 3.5). This allays our main concern about subjects compensating reduced water use per shower with greater frequency of showers each day. Notwithstanding, we are able to reject the null hypothesis of equal treatment effects between the 18L and 28L conditions in phase 1 at the 5% level ($p = 0.040$). This suggests that each shower head in the 28L condition registers 0.284 more showers per day, on average, relative to the 18L condition in phase 1. This is not a huge concern as each shower head registers an average of 3.41 live showers per day in the baseline period, so 0.284 showers (8%) constitute a relatively small fraction. Therefore, we conclude that our treatments only induced adjustments on the intensive margin, so we can focus our attention on it.

Table 3.4: *Effects of moral suasion and real-time feedback on number of showers per day*

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

Notes. This table shows the effects of moral suasion and real-time feedback (based off different shower goals in each phase) on number of *live* showers per day. The results are obtained by estimating a variant of equation (3.1) that includes controls for device and day dummies. For ease of interpretation, we split the estimates from a single regression into 2 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations is reported in the last row. Standard errors clustered at the residence \times floor \times bathroom type level in parentheses.

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

3.3.4 Underlying heterogeneity

Motivated by previous studies showing that high baseline users display larger conservation gains (Allcott, 2011; Ferraro and Price, 2013; Tiefenbeck et al., 2018), we seek to delve into the underlying behavioral mechanisms of our treatments. Formally, we augment our model by including interaction terms with mean baseline use of each shower head, \bar{y}_0 .

$$\begin{aligned}
y_{ith} = & \alpha_i + \lambda_t + \gamma_h + \left(\beta_{MS1}MS_i + \beta_{18L1}18L_i + \beta_{28L1}28L_i \right) \times \text{PHASE1}_{ith} \\
& + \left(\beta_{MS2}MS_i + \beta_{18L2}18L_i + \beta_{28L2}28L_i \right) \times \text{PHASE2}_{ith} \\
& + \left(\gamma_{MS1}MS_i + \gamma_{18L1}18L_i + \gamma_{28L1}28L_i \right) \times \text{PHASE1}_{ith} \times \bar{y}_0 \\
& + \left(\gamma_{MS2}MS_i + \gamma_{18L2}18L_i + \gamma_{28L2}28L_i \right) \times \text{PHASE2}_{ith} \times \bar{y}_0 \\
& + \left(\delta_1\text{PHASE1}_{ith} + \delta_2\text{PHASE2}_{ith} \right) \times \bar{y}_0 + \epsilon_{ith} \tag{3.2}
\end{aligned}$$

In this specification, we include device, date and hour-of-day fixed effects as in equation (3.1). \bar{y}_0 is the mean water use of each shower head in the baseline period, but normalized by subtracting global mean water use (in both residential colleges) in the baseline period. The main coefficients of interest are the γ terms, which tell us how the treatment effects vary with mean baseline use. To account for differential effects by mean baseline use in both phases, we also interacted PHASE1_{ith} and PHASE2_{ith} dummies with \bar{y}_0 respectively. Table 3.5 presents the results.

For the average baseline user, the reported treatment effects are largely similar in each of the experimental conditions. Notwithstanding, it is worth highlighting that there is a clear divergence in treatment effects between the 18L and 28L conditions in phase 2, when the assigned goal was changed to 24L. In particular, we now have sufficient power to reject the null hypothesis of equality at the 1% level ($p = 0.011$), which supports our main finding in Table 3.3.

Zooming in on the γ terms, we find striking differences between the 18L and 28L conditions. In phase 1, while both conditions display similar conservation effects of between 4.7 and 5.2 liters per shower, the underlying treatment dynamics by baseline use stand in stark contrast to each other. More concretely, we observe that there is only a marginal interaction effect, if any, for the 18L condition ($\hat{\gamma}_{18L1} = -0.105; p = 0.073$) but a highly significant effect for the 28L condition ($\hat{\gamma}_{28L1} = -0.327; p = 0.000$). For the latter condition, a one liter increase in mean baseline use increases the treatment effect by approximately 0.33 liters, which amounts to additional gains of 6.3%. We can easily reject the null hypothesis of equality of the interaction terms ($p = 0.002$). This leads us to conclude that while the average treatment effects in phase 1 are similar, the different goals (18L vs. 28L) had in fact induced differing adjustments for residents with different baseline water use behavior. Interestingly, we see that the estimates hold steady in phase 2 when the goal was moved to 24L for both experimental conditions. While there is no significant interaction effect for the 18L condition ($\hat{\gamma}_{18L2} = -0.107; p = 0.114$), we continue to observe a highly significant and quantitatively large effect for the 28L condition ($\hat{\gamma}_{28L2} = -0.388; p = 0.000$). It is also evident that the interaction effects in phase 2 are significantly different from each other ($p = 0.001$).

Table 3.5: *Interaction effects with baseline water use*

Dependent variable:	Water use per shower (liters)	
	PHASE 1 (1)	PHASE 2 (2)
MS \times PHASE	- 0.776 (0.523)	- 0.870 (0.845)
18L \times PHASE	- 4.699*** (0.535)	- 3.885*** (0.877)
28L \times PHASE	- 5.170*** (0.605)	- 6.058*** (0.844)
MS \times PHASE \times \bar{y}_0	0.070 (0.050)	0.167* (0.098)
18L \times PHASE \times \bar{y}_0	- 0.105* (0.058)	- 0.107 (0.067)
28L \times PHASE \times \bar{y}_0	- 0.327*** (0.066)	- 0.388*** (0.086)
PHASE \times \bar{y}_0	- 0.022 (0.032)	- 0.137 (0.085)
Constant	33.908*** (0.244)	
p -value for $\beta_{18L} = \beta_{28L}$	0.455	0.011
p -value for $\beta_{18L1} = \beta_{18L2}$		0.181
p -value for $\beta_{28L1} = \beta_{28L2}$		0.161
p -value for $(\beta_{18L2} - \beta_{18L1}) = (\beta_{28L2} - \beta_{28L1})$		0.003
p -value for $\gamma_{18L} = \gamma_{28L}$	0.002	0.001
p -value for F-test		0.000
Device FEs		Yes
Date FEs		Yes
Hour-of-day FEs		Yes
R^2		0.138
Observations		114627

Notes. This table shows the interaction effects of moral suasion and real-time feedback (with mean baseline use) on water use per shower. The results are obtained by estimating equation (3.2) that includes controls for device, day and hour dummies. For ease of interpretation, we split the estimates from a single regression into 2 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. *live* showers) is reported in the last row. Standard errors clustered at the residence \times floor \times bathroom type level in parentheses.

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

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

3.4 Robustness Checks

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

3.4.1 Sample Selection

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

Notwithstanding, we re-estimate equation (3.1) using the full sample of recorded showers of strictly more than 4 liters. The results are presented in Table 3.7, which offers a useful check for whether our sample selection is representative of the full sample. It is evident that the general result continues to hold true. In particular, we still observe similar average treatment effects between the 18L and 28L conditions in phase 1, followed by a divergence in phase 2. The slight difference is that we now have less power to reject the null hypothesis of equal treatment effects between 18L and 28L conditions in phase 2 ($p = 0.060$). This is not surprising as we are using offline shower data which does not contain information about the actual date and time each shower was taken. To estimate the date and hour-of-day fixed effects here, we could only use the date and time of data upload for each shower event, which is an imperfect proxy at best.¹⁴ Therefore, this exercise provides justification for our sample selection of live showers only.

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

3.4.2 Fraction of time spent under flashing red light

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

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

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

$$y_{ith} = \alpha_i + \lambda_t + \gamma_h + \beta \left(\text{TREAT}_i \times \text{PHASE2}_{ith} \right) + \delta \left(\text{PHASE2}_{ith} \times \bar{y}_0 \right) + \gamma \left(\text{TREAT}_i \times \text{PHASE2}_{it} \times \bar{y}_0 \right) + \epsilon_{ith} \quad (3.3)$$

In this specification, y_{ith} is the outcome variable, i.e. fraction of time spent under flashing red light. TREAT_i is a dummy variable that equals one for the *18L GOAL* (respectively, *28L GOAL*) group for the former (latter) exercise. The *MS* group serves as our “control” and phase 1 is the omitted time period.¹⁵ As in equations (3.1) and (3.2), we include device, date and hour-of-day fixed effects. Therefore, β is the coefficient of interest, which represents the average treatment effect on fraction of time spent under flashing red light for the average baseline user. Table 3.6 presents the results.

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

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

Table 3.6: *Effects of changing goals on time spent under flashing red light*

Dependent variable:	Fraction of time spent under flashing red light	
	18L GOAL (1)	28L GOAL (2)
TREAT \times PHASE ₂	0.025*** (0.008)	- 0.013** (0.007)
PHASE ₂ \times \bar{y}_0	- 0.001 (0.001)	0.001 (0.001)
TREAT \times PHASE ₂ \times \bar{y}_0	0.001 (0.001)	0.000 (0.001)
<i>p</i> -value for F-test	0.000	0.004
Device FEs	Yes	Yes
Date FEs	Yes	Yes
Hour-of-day FEs	Yes	Yes
Mean dependent variable (phase 1)	0.302	0.148
R^2	0.168	0.155
Observations	31054	30069

Notes. This table reports the effects of changing goals on fraction of time spent under flashing red light. The results are obtained by estimating equation (3.3) that includes controls for device, day and hour-of-day dummies. Column 1 reports estimates from the sample of *MS* and *18L GOAL* groups while column 2 reports estimates from the sample of *MS* and *28L GOAL* groups. In both regressions, the *MS* group serves as the “control” and phase 1 is the omitted time period. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. live showers) is reported in the last row. Standard errors clustered at the residence \times floor \times bathroom type level in parentheses.

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

3.4.3 Efficacy of stand-alone feedback lights

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

3.5 Conclusion

Goals are a popular and widely used tool for motivation, effort provision and performance management, both in the private and the public sector. However, if goals can have the properties of reference points, changing them can impose challenges and lead to “side-effects”. Moreover, as behavioral interventions in the form of goal setting and real-time feedback become a valuable tool in the toolkit of resource conservation policies, it becomes even more vital to understand their underlying mechanisms. We run a randomized field experiment in two residential colleges at the National University of Singapore to examine the effect of goals and feedback on performance and the subsequent effect of goal difficulty when goals are changed. One treatment group was assigned a moderate goal in phase 1 (28L), while the other treatment group received a hard goal (18L). In phase 2, both treatment groups got a new (intermediate) goal, which was 24L for both groups.¹⁶ We find that goals alone (in the form of moral suasion) do not have a significant effect on water conservation, however, paired with real-time feedback, there are large and significant conservation effects. Notably, in the first phase, the moderate (28L) and the hard (18L) goal performed equally well on average. However, in phase 2, when both goals are moved to an intermediate 24L goal, differences in performance appear: The former 18L group then performs worse and the former 28L group now shows stronger reactions to the real-time feedback relative to phase 1. This points to the possibility that goal difficulty might not necessary lead to immediate effects and might create lasting effects which shows only once goals are changed.

When looking deeper for the outcomes in phase 1, the fact that the averages are the same masks heterogeneity with respect to baseline use. We find heterogeneity in interaction of reaction towards the real-time feedback (as shown in average treatment effects) with the baseline use between the 18L group and the 28L group. While there is only a marginal interaction effect between baseline use and reaction for the 18L group, we find a highly significant effect for the 28L group between baseline use and treatment effect. Strikingly, this heterogeneity carries over to phase 2, even though the goals are now the same: The 18L group also does not have a significant interaction effect between baseline use and treatment effect in phase 2, while the interaction for the 28L group is highly significant and large. This points to the existence of permanent effects of the initial goal that go beyond our predictions.

Several mechanisms can generate the behavior we observed in our results. It could be either due to loss aversion or fixed penalty (which both give similar predictions) around the goals, or it could be due to psychological disengagement when goals are too difficult. One way to distinguish these mechanisms could be that if the effects are due to reference dependence or the fixed penalty, the outcome would depend on in which direction the goal is moved from the initial “too difficult” goal. For psychological disengagement, it would not matter in which direction one moves from a “too difficult” initial goal as the subjects just stop paying attention to the goal. We cannot make the distinction in this paper since we do not move our subjects from the hard 18L goal to an even harder goal. To further distinguish the underlying behavioral mechanisms could open up promising avenues of future research.

Our findings have important implications for policy makers (and also managers

¹⁶It is important to understand that the treatment groups got different goal difficulties in phase 1 but then the same difficulty in phase 2

and a general audience) since they point out the importance of carefully selecting the optimal level of goal difficulty. In particular, there are two important takeaways from our results:

First, in resource conservation, the role of goals may be particularly important in domains where marginal costs to individuals are low (in our setting, they were zero), as is often the case for water or electricity with certain rental agreements. In these settings, there might be no monetary incentives to save water or electricity and providing such can be costly. While Myers and Souza (2020) found that behavioral channels such as competitiveness, social norms, or moral suasion combined with home energy reports could not increase energy saving efforts without monetary incentives in their setting, our findings suggest that the combination of goals (as moral suasion) and real-time feedback might indeed be a powerful behavioral tool that even works without monetary incentives and when marginal costs of consumption are zero. Our findings are in line with Tiefenbeck et al. (2019) who show that real-time feedback on energy usage during showers leads to higher energy-savings per shower among hotel guests, also in a setting where there are no monetary incentives for energy-saving. This is especially interesting for policy makers since monetary incentives might not always be applicable (as in the case of water and electricity use with certain rental contracts or in hotels) or be costly which make them hard to scale. We show that well designed behavioral interventions such as the combination of goals and real-time feedback might be valuable policy tools for resource conservation which work even in the absence of monetary incentives.

Second, our results suggest that past goals affect current efforts. This is highly relevant since goals are a popular tool for motivation and effort management in the private and public sector alike. The level of difficulty of goals has to be chosen wisely, not only for an optimal current performance but also for an optimal future performance: Too ambitious goals might not only lead to a suboptimal performance but lead to permanent effects which persists even when the goal difficulty is adjusted afterwards. Further research on goal dynamics is needed to shed more light on the underlying interactions and behavioral mechanisms of these findings.

3.A Appendix A: Supplementary figures

3.A.1 Floor plan of the residential colleges



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

Figure 3.8: *Typical floor plan of Cinnamon and Tembusu colleges*

3.A.2 Posters

Figure 3.9: Posters for the Moral Suasion group in each phase



(a) Phase 1

(b) Phase 2

Figure 3.10: Posters for the 18L GOAL group in each phase



(a) Phase 1

(b) Phase 2

Figure 3.11: Posters for the 28L GOAL group in each phase



(a) Phase 1

(b) Phase 2

3.B Appendix B: Supplementary tables

Table 3.7: *Effect of moral suasion and real-time feedback on water use per shower using full sample of recorded showers*

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

Notes. This table shows the effects of moral suasion and real-time feedback (based off different shower goals in each phase) on water use per shower. The results are obtained by re-estimating equation (3.1) using the full sample of recorded showers. We include controls for device, day and hour-of-day dummies. For ease of interpretation, we split the estimates from a single regression into 2 columns, by each phase of the intervention. To reduce the influence of outliers, we drop observations which recorded under 4 liters (inclusive) of water use and top-code using the upper bound of 200 liters. The number of observations (i.e. offline + live showers) is reported in the last row. Standard errors clustered at the residence \times floor \times bathroom type level in parentheses.

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

Chapter 4

Bibliography

- Abrahamse, W., Steg, L., Vlek, C., and Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, 25(3):273–291.
- Abreu, D., Pearce, D., and Stacchetti, E. (1990). Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica: Journal of the Econometric Society*, pages 1041–1063.
- Agarwal, S., Fang, X., Goette, L., Sing, T. F., Staake, T., Tiefenbeck, V., and Wang, D. (2018). The role of goals and real-time feedback in resource conservation: Evidence from a large-scale field experiment. Technical report, National University of Singapore.
- Aghion, P. and Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29.
- Akerlof, G. and Dickens, W. (1982). The economic consequences of cognitive dissonance. *American Economic Review*, 72(3):307–319.
- Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9–10):1082–1095.
- Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.
- Allen, E. J., Dechow, P. M., Pope, D. G., and Wu, G. (2017). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63(6):1657–1672.
- Attari, S. Z., DeKay, M. L., Davidson, C. I., and De Bruin, W. B. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of sciences*, 107(37):16054–16059.
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4):323–370.
- Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

- Becker, L. J. (1978). Joint effect of feedback and goal setting on performance: A field study of residential energy conservation. *Journal of Applied Psychology*, 63(4):428–433.
- Behavioral Insights Team (2013). Applying behavioral insights to organ donation: preliminary results from a randomised controlled trial.
- Bekkers, R. (2006). Traditional and health-related philanthropy: The role of resources and personality. *Social Psychology Quarterly*, 69(4):349–366.
- Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, pages 871–915.
- Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.
- Brookins, P., Goerg, S., and Kube, S. (2017). Self-chosen goals, incentives, and effort. *Unpublished manuscript*.
- Brunnermeier, M. and Parker, J. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.
- Burek, P., Satoh, Y., Fischer, G., Kahil, M., Scherzer, A., Tramberend, S., Nava, L. F., Wada, Y., Eisner, S., Flörke, M., et al. (2016). Water futures and solution-fast track initiative.
- Byrne, D., Goette, L., Martin, L., Schoeb, S., Tiefenbeck, V., and Staake, T. (2018). The behavioral mechanisms of habit formation: evidence from a field experiment. Technical report, University of Melbourne.
- Caille-Brillet, A.-L., Schmidt, K., Watzke, D., and Stander, V. (2015). Bericht zur 2014 Repräsentativstudie „Wissen, Einstellung und Verhalten der Allgemeinbevölkerung zur Organ- und Gewebespende“. *BZgA-Forschungsbericht. Köln: Bundeszentrale für gesundheitliche Aufklärung (BZgA)*.
- Caliendo, M., Fossen, F., and Kritikos, A. (2012). Trust, positive reciprocity, and negative reciprocity: Do these traits impact entrepreneurial dynamics? *Journal of Economic Psychology*, 33(2):394–409.
- Carrillo, J. D. and Mariotti, T. (2000). Strategic ignorance as a self-disciplining device. *Review of Economic Studies*, 67(3):529–544.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree-an open-source platform for laborator, online and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *The American Economic Review*, 99(4):1145–1177.
- Compte, O. and Postlewaite, A. (2012). Belief formation. *Unpublished*.
- Corgnet, B., Gómez-Miñambres, J., and Hernán-Gonzalez, R. (2015). Goal setting and monetary incentives: When large stakes are not enough. *Management Science*, 61(12):2926–2944.

- Crémer, J. (1995). Arm's length relationships. *The Quarterly Journal of Economics*, 110(2):275–295.
- Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine: Journal of the Association of American Medical Colleges*, 78(8):775–780.
- Croson, R. T. (2007). Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry*, 45(2):199–216.
- Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports*, 6:37825.
- Demir, B. and Kumkale, G. T. (2013). Individual differences in willingness to become an organ donor: A decision tree approach to reasoned action. *Personality and Individual Differences*, 55(1):63–69.
- Doerr, J. (2018). *Measure What Matters*. Portfolio; Illustrated Edition (April 24, 2018).
- Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2009). Homo reciprocans: Survey evidence on behavioural outcomes. *Economic Journal*, 119(536):592–612.
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.
- Drucker, P. F. (1954). *The Practice of Management*. Harper, Reissue, Edition 2006.
- Eil, D. and Rao, J. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.
- Enke, B. and Zimmermann, F. (2019). Correlation neglect in belief formation. *Review of Economics Studies*, 86(1):313–332.
- Erez, M. (1977). Feedback: A necessary condition for the goal setting-performance relationship. *Journal of Applied Psychology*, 62(5):624.
- Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3):532–545.
- Farmer, R. E. (1999). *The Macroeconomics of Self-Fulfilling Prophecies*. MIT Press.
- Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, 7(2-3):235–266.
- Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–181.

- Ferraro, P. J. and Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.
- Fisher, G., Kotha, S., and Lahiri, A. (2016). Changing with the times: An integrated view of identity, legitimacy, and new venture life cycles. *Academy of Management Review*, 41(3):383–409.
- Flaxman, S., Goel, S., and Rao, J. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, pages 298–320.
- Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *Quarterly Journal of Economics*, 126(4):1799–1839.
- Goerg, S. J., Kube, S., and Radbruch, J. (2019). The effectiveness of incentive schemes in the presence of implicit effort costs. *Management Science*, 65(9):4063–4078.
- Goette, L., Leong, C., and Qian, N. (2019). Motivating household water conservation: A field experiment in Singapore. *PloS one*, 14(3).
- Gómez-Miñambres, J. (2012). Motivation through goal setting. *Journal of Economic Psychology*, 33(6):1223–1239.
- Grossman, Z. and Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510–524.
- Grove, A. S. (1983). *High Output Management*. Vintage; 2nd Edition (August 29, 1995).
- Harding, M. and Hsiaw, A. (2014). Goal setting and energy conservation. *Journal of Economic Behavior & Organization*, 107:209–227.
- Heath, C., Larrick, R., and Wu, G. (1999). Goals as Reference Points. *Cognitive Psychology*, 38:79–107.
- Herr, A. and Normann, H.-T. (2016). Organ donation in the lab: Preferences and votes on the priority rule. *Journal of Economic Behavior & Organization*, 131:139–149.
- Herr, A. and Normann, H.-T. (2019). How much priority bonus should be given to registered organ donors? An experimental analysis. *Journal of Economic Behavior & Organization*, 158:367–378.
- Herweg, F., Müller, D., and Weinschenk, P. (2010). Binary payment schemes: Moral hazard and loss aversion. *American Economic Review*, 100(5):2451–77.
- Hill, E. M. (2016). Posthumous organ donation attitudes, intentions to donate, and organ donor status: Examining the role of the big five personality dimensions and altruism. *Personality and Individual Differences*, 88:182–186.
- Iyengar, S. and Hahn, K. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39.

- Jehiel, P. and Steiner, J. (2019). Selective sampling with information-storage constraints, and resulting behavioral biases. *The Economic Journal*.
- Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–291.
- Kata, A. (2010). A postmodern pandora’s box: anti-vaccination misinformation on the internet. *Vaccine*, 28(7):1709–1716.
- Kennerley, M. and Neely, A. (2003). Measuring performance in a changing business environment. *International Journal of Operations & Production Management*.
- Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800):829–832.
- Kőszegi, B. (2003). Health anxiety and patient behavior. *Journal of Health Economics*, 22(6):1073–1084.
- Langenbach, B. P., Berger, S., Baumgartner, T., and Knoch, D. (2019). Cognitive resources moderate the relationship between pro-environmental attitudes and green behavior. *Environment and Behavior*.
- Latham, G. P. and Locke, E. A. (2006). Enhancing the benefits and overcoming the pitfalls of goal setting. *Organizational Dynamics*, 35(4):332–340.
- Lavee, J., Ashkenazi, T., Gurman, G., and Steinberg, D. (2010). A new law for allocation of donor organs in Israel. *Lancet*, 375(9720):1131.
- Leung, B. T. K. (2020). Limited cognitive ability and selective information processing. *Games and Economic Behavior*, 120:345–369.
- Levin, J. (2003). Relational incentive contracts. *American Economic Review*, 93(3):835–857.
- Liang, Y. (2019). Learning from unknown information sources. *Working Paper*.
- Locke, E. A. and Latham, G. P. (1990). *A theory of goal setting & task performance*. Prentice-Hall, Inc.
- Locke, E. A. and Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705.
- Locke, E. A. and Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, 15(5):265–268.
- Martinez, J. M., Lopez, J. S., Martin, A., Martin, M. J., Scandroglio, B., and Martin, J. M. (2001). Organ donation and family decision-making within the spanish donation system. *Social Science & Medicine*, 53(4):405–421.

- McCoy, J., Rahman, T., and Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1):16–42.
- Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2014). Managing self-confidence. *Working Paper*.
- Mocan, N. and Tekin, E. (2007). The determinants of the willingness to donate an organ among young adults: evidence from the united states and the european union. *Social Science & Medicine*, 65(12):2527–2538.
- Morgan, S. and Miller, J. (2002). Communicating about gifts of life: The effect of knowledge, attitudes, and altruism on behavior and behavioral intentions regarding organ donation. *Journal of Applied Communication Research*, 30(2):163–178.
- Myers, E. and Souza, M. (2020). Social comparison nudges without monetary incentives: Evidence from home energy reports. *Journal of Environmental Economics and Management*, page 102315.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175.
- Nijkamp, M. D., Hollestelle, M. L., Zeegers, M. P., van den Borne, B., and Reubsæet, A. (2008). To be (come) or not to be (come) an organ donor, that’s the question: A meta-analysis of determinant and intervention studies. *Health Psychology Review*, 2(1):20–40.
- Ordóñez, L. D., Schweitzer, M. E., Galinsky, A. D., and Bazerman, M. H. (2009). Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives*, 23(1):6–16.
- Oyer, P. (2000). A theory of sales quotas with limited liability and rent sharing. *Journal of Labor Economics*, 18(3):405–426.
- Pang, D., Bleetman, A., Bleetman, D., and Wynne, M. (2017). The foreign body that never was: the effects of confirmation bias. *British Journal of Hospital Medicine*, 78(6):350–351.
- Perugini, M., Gallucci, M., Presaghi, F., and Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality*, 17(4):251–283.
- Radecki, C. M. and Jaccard, J. (1997). Psychological aspects of organ donation: a critical review and synthesis of individual and next-of-kin donation decisions. *Health Psychology*, 16(2):183.
- Redelmeier, D. A. and Woodfine, J. D. (2013). Deceased organ donation and the Nicholas effect. *Transplantation*, 96(11):e82–e84.
- Richter, D., Metzging, M., Weinhardt, M., and Schupp, J. (2013). SOEP scales manual. *SOEP Survey Papers*.
- Roach, K. (2010). Wrongful convictions: Adversarial and inquisitorial themes. *North Carolina Journal of International Law and Commercial Regulation*, 35(2):387.

- Rönnerstrand, B. and Sundell, K. A. (2015). Trust, reciprocity and collective action to fight antibiotic resistance. An experimental approach. *Social Science & Medicine*, 142:249–255.
- Schmidt, K., Watzke, D., and Stander, V. (2013). "Einstellung, Wissen und Verhalten der Allgemeinbevölkerung zur Organ- und Gewebespende." Zusammenfassung der wichtigsten Ergebnisse". Köln: Bundeszentrale für gesundheitliche Aufklärung.
- Schmidt, K., Watzke, D., and Stander, V. (2014). "Wissen, Einstellung und Verhalten der Allgemeinbevölkerung zur Organ- und Gewebespende." Zusammenfassung der wichtigsten Ergebnisse der Repräsentativbefragung 2013. Köln: Bundeszentrale für gesundheitliche Aufklärung.
- Sehgal, N. K., Scallan, C., Sullivan, C., Cedeño, M., Pencak, J., Kirkland, J., Scott, K., and Thornton, J. D. (2016). The relationship between verified organ donor designation and patient demographic and medical characteristics. *American Journal of Transplantation*, 16(4):1294–1297.
- Shacham, E., Loux, T., Barnidge, E. K., Lew, D., and Pappaterra, L. (2018). Determinants of organ donation registration. *American Journal of Transplantation*, 18(11):2798–2803.
- Siegel, J. T., Tan, C. N., Rosenberg, B. D., Navarro, M. A., Thomson, A. L., Lyrantzis, E. A., Alvaro, E. M., and Jones, N. D. (2016). Anger, frustration, boredom and the department of motor vehicles: Can negative emotions impede organ donor registration? *Social Science & Medicine*, 153:174–181.
- Speier, C., Valacich, J. S., and Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, 30(2):337–360.
- Stigler, G. J. and Becker, G. S. (1977). De Gustibus Non Est Disputandum. *American Economic Review*, 67(2):76–90.
- Stoler, A., Kessler, J. B., Ashkenazi, T., Roth, A. E., and Lavee, J. (2017). Incentivizing organ donor registrations with organ allocation priority. *Health Economics*, 26(4):500–510.
- Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. (2018). Overcoming salience bias: how real-time feedback fosters resource conservation. *Management Science*, 64(3):1458–1476.
- Tiefenbeck, V., Wörner, A., Schöb, S., Fleisch, E., and Staake, T. (2019). Real-time feedback promotes energy conservation in the absence of volunteer selection bias and monetary incentives. *Nature Energy*, 4(1):35–41.
- Tversky, A. and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics*, 106(4):1039–1061.
- van Andel, C. E., Tybur, J. M., and Van Lange, P. A. (2016). Donor registration, college major, and prosociality: Differences among students of economics, medicine and psychology. *Personality and Individual Differences*, 94:277–283.

- Wakefield, C. E., Watts, K. J., Homewood, J., Meiser, B., and Siminoff, L. A. (2010). Attitudes toward organ donation and donor behavior: a review of the international literature. *Progress in Transplantation*, 20(4):380–391.
- Walkowitz, G. (2019). On the validity of probabilistic (and cost-saving) incentives in dictator games: A systematic test. MPRA Paper No. 91541, Technical University of Munich, Munich, Germany.
- Wilson, A. (2014). Bounded memory and biases in information processing. *Econometrica*, 82(6):2257–2294.
- World Economic Forum (2020). The Global Risks Report 2020 - 15th Edition.
- Wu, G., Heath, C., and Larrick, R. (2008). A prospect theory model of goal behavior. *Unpublished manuscript*.
- Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., and Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, 109(50):20364–20368.