Essays in Experimental Economics

Thesis by
Polina Detkova

In Partial Fulfillment of the Requirements for the
Degree of
PhD in Social Science

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 12, 2025

Polina Detkova
ORCID: 0000-0002-4716-2758

ACKNOWLEDGEMENTS

First, I want to thank my advisor, Marina Agranov, for her guidance, insight, and unwavering support throughout my graduate studies. I am also deeply grateful to my committee members—Charles Sprenger, Kirby Nielsen, and Thomas Palfrey—for their thoughtful feedback, critical questions, and generous engagement with my work. Their expertise has deepened both my projects and my development as a scholar beyond anything I could have expected.

I am especially thankful to my parents, Viacheslav Detkov and Liudmila Detkova, for their love, encouragement, and endless patience. I also owe special thanks to my fellow graduate students Peter Doe and Egor Stoyan for their steadfast academic and personal support throughout this journey. Finally, I want to thank the broader community of faculty, graduate students, and administrative staff in the HSS department, whose support, feedback, and companionship have enriched this experience in countless ways.

ABSTRACT

This dissertation consists of three essays that use lab and online experiments to investigate how individuals make decisions under uncertainty, in social contexts, and when forming beliefs about others. Each essay introduces a distinct setting, but all share a common goal, which is to improve our understanding of human decision-making.

Chapter 1 examines commitment contracts. Their high rates of failure raise concerns since individuals may end up worse off than if they had never committed. We investigate whether some of these failures are actually anticipated, with individuals recognizing that future uncertainty might make failing the contract the best option upon some realizations of uncertainty. We refer to this behavior as *planning for the possibility of failure*. This approach is different from the usual interpretation of failures, which we call *failing to plan*, as it attributes failures to take-up mistakes. To study whether individuals plan for the possibility of failure, we conducted a controlled lab experiment designed to detect patterns of such planning. Our findings indicate that about one-third of all commitment choices can be attributed to this kind of foresight. This suggests that planning for failure is common, and that high failure rates are not necessarily driven by mistaken commitments. Thus, they do not by themselves call into question the value of commitment contracts.

The second essay studies the decision to ask for help—a behavior that can be critical in addressing information asymmetries but is often avoided. In an online experiment, we find that making potential helpers even minimally identifiable (e.g., through an uninformative ID number) significantly increases the likelihood of asking. Belief data suggest that this effect stems from shifts in how individuals weigh expected payoffs and other factors (particularly social ones) when deciding whether to ask.

The third essay explores how people expect others to update their beliefs upon receiving new information. We find that when two individuals have different priors, people expect others' beliefs to move toward their own prior upon receiving new information. Although this result is consistent with the theoretical predictions for Bayesian agents, we find no support for the precision of information affecting the magnitude of the shift in the way the theory predicts. We find that this effect occurs not only due to under-updating of one's own beliefs but also due to recognition of under-updating by others.

PUBLISHED CONTENT AND CONTRIBUTIONS

Agranov, M., & Detkova, P. (2024). *Beliefs of Others: An Experiment* (Working paper). https://agranov.caltech.edu/documents/29592/paperIVP.pdf
D.P. participated in creating the experiment design and analyzing the data.

Detkova, P. (2025). *The Identified Helper Effect on the Frequency of Asks* (Working paper).

Detkova, P., & Stoyan, E. (2025). *Failing to Plan or Planning to Fail? A Study on Commitment Failure* (Working paper).
D.P. participated in the conception of the project, theoretical modeling, creating the experiment design, and conducted the data analysis and wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

INTRODUCTION

This dissertation presents three essays in experimental economics, each based on either a laboratory or an online experiment. Experimental methods offer a uniquely powerful way to test economic theories and uncover behavioral mechanisms that are otherwise difficult to observe in the field. By creating controlled environments where key variables can be isolated and systematically varied, experiments allow us to directly measure how people respond to uncertainty, incentives, and social context. Together, these essays highlight the richness of experimental economics as a tool for uncovering patterns in human decision-making.

Chapter 1 studies commitment contracts. They are designed for time-inconsistent agents, who are torn between an immediate pleasure and pursuing a long-term goal, such as watching TV vs. going to the gym. Thus, one can predict it would be hard for them to go to the gym tomorrow, so they can promise $10 to their friend if they skip it. Despite the threat of a penalty designed to keep people consistent, studies on commitment contracts report high failure rates and the corresponding frequent loss of money. The standard interpretation in the literature says that failures happen due to *failing to plan* the contracts. For example, people could underestimate the size of the penalty required (Bai et al., 2021; Carrera et al., 2022; Heidhues and Kőszegi, 2009).

We study whether individuals expecting future uncertainty sometimes make it optimal to fail their commitment contracts under some uncertainty realizations. We call this *planning for the possibility of failure*. The state of the world, in which the participants are to do the action, is inherently uncertain, as the action occurs in the future. For people to commit, there should be some states of the world, in which the contract encourages them to do the action. But there is no need for *all* states of the world to be like that. So long as commitment is beneficial in expectation, the ex ante possibility of failure does not matter.

In our lab experiment designed to explore whether people plan for the possibility of failure, participants come to the lab twice. In the experiment, the participants had the opportunity to transcribe lines of Greek letters. On Day 1, they could commit to transcribing a certain number of them, and one week later, on Day 2, they could conduct the transcriptions. In the experiment, we can control the cost of action and make it, by chance, *low* or *high*. Thus, on Day 1, the participants know that they will

face, by chance, low or high costs when the time comes to do the action. However, they only have an opportunity to commit on Day 1, so the commitment will have to apply regardless of the cost realization. As the costs are either low or high, planning for the possibility of failure becomes equivalent to planning to fail upon the high realization of costs.

Our participants took up around half of all commitment contracts we offered to them, while about 70% of participants took up at least one option available. If people do not plan for the possibility of failure, all contracts for uncertain costs would provide enough motivation under the high realization of costs. However, more than a third of the contracts in the data lack the necessary motivation. We conclude that planning for the possibility of failure is extensive. It has many implications, and one suggests that the money people lose due to commitment is not necessarily a by-product of their mistakes.

Chapter 2 suggests and studies a practical tool for increasing the frequency of asking for help. People in need often choose not to ask for help, despite the fact that asking can be critical for receiving it. Prior research highlights that asking increases giving (Andreoni and Rao, 2011), yet many refrain from asking due to psychological and social barriers (Bohns, 2016; Jaroszewicz et al., 2022; Lee, 1997). Our study contributes to the effort of encouraging asks by testing a simple, scalable intervention that makes potential helpers only marginally more identifiable—assigning them uninformative ID numbers.

In our Prolific experiment, participants were randomly assigned to be askers or helpers. Helpers received higher initial payments, while askers could request help by asking for a transfer from a helper. In the Treatment ID condition, askers were shown the uninformative ID of their potential helper; in the Treatment No ID condition, they were not. We find that displaying even a weak form of identifiability significantly increases asking rates—from 67% to 76.5%. Moreover, the role of beliefs differs sharply between treatments: in Treatment ID, askers' expected payoff predicts asking, while in Treatment No ID, other factors—such as expectations about others' asking behavior—play a stronger role.

These findings advance the literature on help-seeking by offering a practical tool for increasing asks and providing a novel explanation for identifiability effects. We show that even minimal identifiability can change behavior and that beliefs about others matter more when identifiability is absent. Our study complements existing psychological and neurological explanations of identifiability effects and

demonstrates that these phenomena extend beyond laboratory settings to online platforms like Prolific. The simplicity of the intervention suggests potential for broad application in charitable, organizational, and policy contexts.

Chapter 3 explores how people expect others to update their beliefs. While the belief-updating process itself has been extensively studied (see, for example, Benjamin, 2019, for an overview), much less is known about how individuals think others revise their beliefs in response to new information. From a theoretical perspective, Kartik et al., 2021 examine this question for Bayesian agents. Let Bob be the person who receives a signal about the true state of the world, and let Anne be the person who forms an expectation about Bob's posterior. If Anne and Bob have different non-degenerate priors over a binary state of the world, Kartik et al., 2021 show that: (1) Anne expects Bob's posterior to shift closer to her own prior; and (2) the magnitude of this shift increases with the precision of the signal.

We use a Prolific experiment to test this theory. Unlike most studies on belief updating, we elicit individual home-grown beliefs about factual statements,[1] rather than constructing an artificial environment. We elicit individuals' own beliefs, their beliefs about others' posteriors conditional on the signal realization, and their beliefs about the average posterior of other participants.

Although we find support for the first part of the Bayesian prediction, there is little evidence that signal precision affects Anne's beliefs about the average posterior of others. This result has important implications for Bayesian-based models of belief updating about others. For example, in Kartik et al., 2021, the precision effect plays a central role in predicting that more precise tests may deter some individuals from taking them. Our results suggest that such concerns may be overstated: the potential reduction in participation due to increased precision appears limited compared to the informational gains from more precise testing.

References

Andreoni, J., & Rao, J. M. (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics*, *95*(7-8), 513–520. https://doi.org/10.1016/j.jpubeco.2010.12.008 (cit. on p. 2).

[1] For example, we asked about the chance that the statement "RHINO HORN IS MADE UP OF KERATIN—THE SAME PROTEIN WHICH FORMS THE BASIS OF OUR HAIR AND NAILS" is correct.

Bai, L., Handel, B., Miguel, E., & Rao, G. (2021). Self-Control and Demand for Preventive Health: Evidence from Hypertension in India. *The Review of Economics and Statistics*, *103*(5), 835–856. https://doi.org/10.1162/rest_a_00938 (cit. on p. 1).

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations*, *1*, 69–186 (cit. on p. 3).

Bohns, V. K. (2016). (Mis)Understanding Our Influence Over Others: A Review of the Underestimation-of-Compliance Effect. *Current Directions in Psychological Science*, *25*(2), 119–123. https://doi.org/10.1177/0963721415628011 (cit. on p. 2).

Carrera, M., Royer, H., Stehr, M., Sydnor, J., & Taubinsky, D. (2022). Who Chooses Commitment? Evidence and Welfare Implications. *The Review of Economic Studies*, *89*(3), 1205–1244. https://doi.org/10.1093/restud/rdab056 (cit. on p. 1).

Heidhues, P., & Kőszegi, B. (2009). Futile Attempts at Self-Control. *Journal of the European Economic Association*, *7*(2-3), 423–434. https://doi.org/10.1162/JEEA.2009.7.2-3.423 (cit. on p. 1).

Jaroszewicz, A., Loewenstein, G., Canavan, K., Martin, J., & Tevar, A. (2022). *The Psychological Pain of Asking for Live Kidney Donations* (Working paper). (Cit. on p. 2).

Kartik, N., Lee, F. X., & Suen, W. (2021). Information validates the prior: A theorem on bayesian updating and applications. *American Economic Review: Insights*, *3(2)*, 165–182 (cit. on p. 3).

Lee, F. (1997). When the Going Gets Tough, Do the Tough Ask for Help? Help Seeking and Power Motivation in Organizations. *Organizational Behavior and Human Decision Processes*, *72*(3), 336–363. https://doi.org/10.1006/obhd.1997.2746 (cit. on p. 2).

Chapter 1

FAILING TO PLAN OR PLANNING TO FAIL: A STUDY ON COMMITMENT FAILURE

1 INTRODUCTION

People regularly fail to achieve long-term goals, which is often attributed to time inconsistency (DellaVigna and Malmendier, 2004; Laibson, 1997; O'Donoghue and Rabin, 1999). For example, people make New Year's resolutions, which they often cannot maintain even for a week (Norcross and Vangarelli, 1988). Time inconsistency provides a specific mechanism for such failures: temptation and self-control issues. Time-inconsistent individuals may exhibit a disproportionate preference for immediate consumption, leading them to deviate from long-term objectives.

Commitment contracts are designed to prevent such deviations. They discourage people from drinking alcohol (Schilbach, 2019), smoking cigarettes (Chaloupka et al., 2019, December; Giné et al., 2010; Halpern et al., 2015), skipping doctor appointments (Bai et al., 2021) and gym sessions (Carrera et al., 2022; Royer et al., 2015). Yet, these studies regularly show that at least 30% of participants fail to follow through with their commitments as summarized by John, 2020. The most widely discussed explanation is underestimating one's present bias: individuals commit, but the incentives provided by the contract turn out to be too weak to change their behavior (Bai et al., 2021; Heidhues and Kőszegi, 2009). Mean-zero mistakes in evaluating the benefits of a contract are another possible source of failures (Carrera et al., 2022). From both of these perspectives, the failures are unplanned: an individual takes up a commitment contract while aiming to follow through.

In this study, we consider an alternative possibility: when individuals face uncertainty, can they deliberately choose the contracts they can, by chance, fail? We call this phenomenon the *planning for the possibility of failure* motive. For example, an individual may promise her friend to pay him $10 if she misses a gym session the following day. We can expect this commitment to force her to go to the gym on a regular day, but if she feels sick, she might pay $10 and miss the session. If, instead, she promised to pay $100 to her friend, we can expect her to go to the gym despite it being a suboptimal action. That is why she might want a contract with a

$10 penalty while being aware that, with some chance, she may fail it.

By studying if people can plan for the possibility of failure, we take the first step toward assessing the empirical relevance of planned failures, as this process can be challenging. For example, asking an individual about her chance of failing a commitment contract is far from innocuous. If the individual is asked before committing, the question itself can interfere with her choice. If asked after deciding whether to commit, her decision on commitment take-up can affect her response.[1] Thus, a committed individual might overestimate her chance of following through to ex-post validate her decision. Meanwhile, evidence from other economic fields suggests that people might fail to incorporate future uncertainty into their decisions (Kueng, 2018; Leary and Wang, 2016; Shapira and Venezia, 1999), which makes our research question far from trivial.

To answer our research question, we create uncertain costs in a lab experiment and allow our participants to design their own commitment contracts. Our participants engage in Greek letter transcriptions, which are regularly chosen for commitment studies in laboratory settings (Augenblick, Niederle, and Sprenger, 2015; Ek and Samahita, 2023). We introduce uncertain costs by randomly assigning Short or Long tasks to the participants. These tasks are paid equally per amount completed; however, it takes longer to complete one Long task. Subjects can choose a commitment on the minimal amount of tasks to be completed in the future. This commitment consists of a number of tasks and an amount of money: if a participant commits and then does fewer tasks than the contract states, she loses the money. Notably, the participants who want to commit can select the number of tasks and the amounts of money across a wide range of positive numbers.

We focus on the contracts that state the number of tasks and the amount of money to *always* apply: no matter if the participants, by chance, get Short or Long tasks. We compare these contracts to the ones that apply contingently to getting Long tasks. We assume that the contract for Long tasks motivates her to do precisely the number of Long tasks required by this contract. Therefore, if a contract that always applies is more demanding than the one contingent on getting Long tasks,[2], we conclude that this contract is planned to be failed upon getting Long tasks. We call such contracts *partial*. Our primary goal is to estimate the proportion of partial contracts, which would be the lower bound of all contracts planned to be failed.

[1] The 'illusion of control,' described by Langer, 1975, can be one of the reasons.

[2] I.e., it either is supposed to motivate doing more tasks using the same penalty, or uses a lower penalty to motivate doing at least as many tasks.

We conduct our experiment with 76 subjects drawn from the UCSD undergraduate population. We estimate the share of partial contracts to be 30–40%. We test the validity of our partial contracts measure based on the theoretical result connecting the proportion of contracts planned to be failed and the costs of doing the tasks. We find a positive relationship between the proportion of partial contracts and the costs of doing the tasks, which is aligned with the model's prediction.

The primary implication of our study is that commitment contracts as a tool should not be dismissed based on extensive failures. These failures led a few recent studies to demonstrate a decrease in welfare after introducing commitment options (Bai et al., 2021; Carrera et al., 2022; John, 2020). Yet, as we find individuals capable of planning for the possibility of failure, the source of failures requires examination beyond that bound to specific modeling approaches. In the paper, we argue that the planning for the possibility of failure motive can explain many failures in previous studies. Even failure rates higher than 30% have the potential to stem entirely from planning (in the studies by Carrera et al., 2022; John, 2020; Royer et al., 2015, based on our assessment).

We collect empirical evidence that allows us to make other contributions to studies on commitment contracts. First, we observe extensive take-up of commitment options with very low penalties for failure (below $1). One interpretation is that the participants are mostly unaware of their time inconsistency. This result is usually obtained by methods not based on commitment choices (Augenblick and Rabin, 2019; Fedyk, 2016). Meanwhile, commitment-based estimations suggest that people are primarily aware of their time inconsistency (Bai et al., 2021; Carrera et al., 2022; Chaloupka et al., 2019, December). Therefore, the lack of extremely low penalties in previous studies can be one of the reasons for this gap. Second, we observe extensive commitment take-up just thirty minutes before doing the tasks. On the one hand, it can be analyzed as any take-up made in advance. On the other hand, this result can be considered evidence of commitment take-up for reasons other than addressing one's time inconsistency (Bonein and Denant-Boèmont, 2015; Kaur et al., 2015; Schilbach, 2019).

The policy implications of our study are not limited to the potential use of commitment contracts despite high failure rates. We find that demand for *some* commitment is mostly determined by individual characteristics rather than characteristics of a specific contract. It suggests an opportunity to offer corrections to address individuals' mistakes. For example, the participants willing to commit can be explained

by the tendency people have to underestimate their time inconsistency, and can be offered a more restrictive contract instead of their own choice.

The paper proceeds as follows. Section 2 reviews the related studies. Section 3 shows the theoretical model that develops our identification strategy. Section 4 describes the design of our lab experiment and the sample. Section 5 provides the results of our experiment, and Section 6 discusses them. Section 7 concludes.

2 RELATED LITERATURE

Commitment contracts are designed to help time-inconsistent individuals achieve long-term goals. Many of these contracts consist of an action and a monetary penalty: if an individual fails to perform the action in the future, they must pay the penalty. We refer to these contracts as *penalty-based*.[3] Unlike the contracts that reward people for their behavior, penalty-based contracts are self-financed and do not require external funding. These contracts have been shown to discourage engagement in undesirable actions such as drinking alcohol (Schilbach, 2019), smoking cigarettes (Giné et al., 2010; Halpern et al., 2015), skipping doctor appointments (Bai et al., 2021), and missing the gym (Carrera et al., 2022; Royer et al., 2015). Other contracts have helped many individuals save money (John, 2020), make healthier food choices (Schwartz et al., 2014), and achieve other goals (Burger and Lynham, 2010; Exley and Naecker, 2017; Kaur et al., 2015).

Commitment failures attract attention due to their surprisingly large scale.[4] There is no clear reason for regular failures since participants voluntarily take up commitment contracts in the aforementioned studies. Therefore, we can expect them to follow through with the contracts and keep the money. The naivete of participants has long been considered the primary reason for failures in penalty-based contracts (Bai et al., 2021; Heidhues and Kőszegi, 2009; John, 2020). Intuitively, participants may underestimate the large penalty required to motivate the desirable action and choose a commitment contract with too small a penalty. Then, when the time comes to perform the action, they not only fail to do so but also incur the penalty. Carrera et al., 2022 suggests that stochastic valuation errors may also contribute to

[3] Other commitment contracts might reward people for their behavior (Acland and Levy, 2015; Aksoy et al., 2023; Allcott, Kim, et al., 2021; Avery et al., 2019; Chaloupka et al., 2019, December; Severine Toussaert, 2019) or impose restrictions on it (Allcott, Gentzkow, and Song, 2022; Brune et al., 2021; Casaburi and Macchiavello, 2019; John, 2020).

[4] Column 6 of Table 1.3 provides examples of the proportion of penalty-based commitment contracts that are failed. John, 2020 summarizes failures across different types of contracts.

this issue. They model the utility of a commitment contract as random, where a contract-specific term amplifies a mean-zero error term. Despite the error being symmetric around zero, it can considerably contribute to mistaken contract take-up. Overall, different reasons for take-up mistakes are currently explored as the main explanation for extensive failures.

While we are the first to empirically demonstrate that individuals can plan for occasional failures, we are not the first to introduce stochastic costs into the analysis. Laibson, 2015 examined their impact theoretically by analyzing take-up and failures for a uniform distribution of the costs of performing an action. John, 2020 introduces stochastic shocks to monthly income but finds this type of shocks incapable of explaining the failures in her study. Carrera et al., 2022 introduced a stochastic distribution of costs for going to the gym but found that the change in welfare was overall negative, which is incompatible with failures being anticipated by the participants. At the same time, Allcott, Kim, et al., 2021 emphasize that it is the stochastic repayment cost shocks that allow their model of payday loans to be compatible with the data. They found that access to payday loans increased participants' welfare across many model specifications.

So far, the possibility that people anticipate commitment failures has been introduced in theoretical modeling, but it has been unclear whether individuals adjust their commitment contract take-up as we expect them to. Notably, individuals have only limited ability to adjust their consumption (Kueng, 2018) and payday loans (Leary and Wang, 2016) even in response to deterministic changes in their future income. Further, Shapira and Venezia, 1999 find that, given the role of insurance sellers, individuals cannot do screening between 'careful' and 'negligent' people using contract prices. In addition, participants can neglect lower realizations of costs altogether if they underestimate their time-inconsistency and believe that their actions do not need to be incentivized if the costs are low.

The take-up data in our study suggests that many participants might commit for reasons other than to manage one's time inconsistency. Previous studies discuss such reasons as signaling others (Exley and Naecker, 2017), peer pressure (Bonein and Denant-Boèmont, 2015), and social pressure (Kaur et al., 2015; Schilbach, 2019). Notably, these reasons are different from the factors that interfere in the process of managing one's time-inconsistency, such as valuation errors (Carrera et al., 2022) or costly experimentation (Kaur et al., 2015).

Estimates of the perceived present bias parameter $\widehat{\beta}$ vary between different studies.

Estimations based on commitment choices (Bai et al., 2021; Carrera et al., 2022; Chaloupka et al., 2019, December) suggest it to be significantly below 1, which is the naivete benchmark for time-inconsistent agents. Meanwhile, other methods find participants naive about their present bias (Augenblick and Rabin, 2019; Fedyk, 2016). In our study, we give the participants an opportunity to commit under very low penalties. The small penalties that participants choose in our study fully correspond to the estimates of the perceived present bias parameter being close to 1: individuals can believe that even a negligible penalty would be enough to change their behavior. We merely suggest the lack of flexibility in commitment contracts as a reason for discrepancy. This is because there are many differences between the aforementioned studies; for example, the estimates of commitment contracts were conducted in the field, while other studies were conducted in the lab.

Our approach to stochastic costs and anticipation of failure connects our study to other economic fields. In a standard principal-agent framework, we can consider an agent's current self as a principal and her future self as an agent. Then, a commitment contract that allows the individuals to plan for occasional failures becomes a screening device for agents with different uncertainty realizations. Shapira and Venezia, 1999 gave their experiment participants an opportunity to act as insurance sellers that can offer two types of contracts: with full coverage and a deductible. They found that feedback and instructions were required before participants could set the prices to these contracts so that they could screen between 'careful' and 'negligent' potential buyers. The study by Posey and Yavas, 2007 supports this idea by providing evidence that creating an experimental insurance market leads to the predictions of screening equilibrium over time.

Our results are also connected to a broad discussion on Keynesian decision-makers: can people adjust their decisions to uncertainty in the future? Individuals have only limited ability to adjust their consumption Kueng, 2018 and payday loans Leary and Wang, 2016 even in response to deterministic changes in their future income. We find our participants capable of adjusting their commitment contracts to the uncertainty in the future. This makes our results in line with the literature on the preferences for commitment and flexibility, which assume that individuals can plan ahead (Afzal et al., 2019, May; Amador et al., 2006; Galperti, 2015; Séverine Toussaert, 2018).

3 MODEL

In our experiment, we exogenously introduce uncertain costs, which are realized *after* individuals decide whether to commit or not. In this section, we build a theoretical framework for identifying the contracts that are planned to be sometimes failed. We consider only two realizations of cost functions, which correspond to our experiment design. However, the straightforward intuition behind our model makes it adaptable to other setups or distributions of costs.

3.1 Setup

We consider a two-period model for a $(\beta, \widehat{\beta}, \delta)$ agent (O'Donoghue and Rabin, 2001). The agent has an opportunity to earn money by completing Greek-letter transcription tasks. A participant can commit to completing a certain number of tasks. Specifically, she can take up a penalty-based commitment in period 1. In period 2, the agent engages in the tasks and receives payment. Since there is no time difference between the engagement in the tasks and payment—neither in the model nor in our experiment—we omit the exponential discounting parameter δ from consideration by setting it equal to 1. Further, we assume that the present bias effect on the monetary amounts is negligible compared to that on consumption (Augenblick, Niederle, and Sprenger, 2015; Cerrone et al., 2023; Imai et al., 2021).

We introduce uncertain costs by making the tasks, by chance, either Short (S) or Long (L); each Short task contains fewer Greek letters than a Long one. The agent gets all her tasks to be Short with chance $0 < p < 1$, and all her tasks to be Long with chance $1 - p$. She learns the length of her tasks at the beginning of period 2. Therefore, the agent has the opportunity to commit only while the costs are uncertain. Regardless of the length of her tasks, the agent receives a piece-rate wage w for each task she completes. We assume that the agent maximizes her expected consumption flow and that her consumption is quasilinear in money.

In period 1, the agent prefers her future self to complete $t^\star(\gamma)$ tasks in period 2 if she receives tasks of the difficulty γ:

$$t^\star(\gamma) = \arg\max_t \; wt - C(t, \gamma), \tag{1.1}$$

where the function $C(t, \gamma)$ referees to the costs of completing t tasks of the difficulty γ. We denote longer tasks with larger values of γ. We impose a few assumptions on the Cost function $C(t, \gamma)$.

(I) Doing 0 tasks has no cost, regardless of task complexity. $C(0,\gamma) = 0\ \forall\gamma$.

(II) Costs are increasing and convex w.r.t. the number of tasks t. $\frac{\partial C(t,\gamma)}{\partial t} > 0$ and $\frac{\partial^2 C(t,\gamma)}{\partial t^2} > 0$.

(III) She gets tired faster from an additional task when doing more complex tasks $\frac{\partial^2 C(t,\gamma)}{\partial t\,\partial\gamma} > 0$.

(IV) Costs increase faster when doing more complex tasks $\frac{\partial^3 C(t,\gamma)}{\partial t^2\,\partial\gamma} > 0$.

In period 2, the time inconsistency of the agent may alter her decision regarding the optimal effort compared to period 1. Using the present bias parameter β, we say that in period 2, the agent prefers to complete $t_2^\star(\gamma)$ tasks if she receives tasks of the complexity γ:

$$t_2^\star(\gamma) = \arg\max_t \ wt - \frac{1}{\beta}C(t,\gamma). \tag{1.2}$$

If the agent is time inconsistent, then $\beta \neq 1$. We focus on a present-biased agent with $\beta < 1$. Parameter $\widehat{\beta}$ captures the agent's perceptions of her present bias. She might be naive about her present bias ($\widehat{\beta} = 1$), be partially sophisticated ($\beta < \widehat{\beta} < 1$), or fully sophisticated ($\beta = \widehat{\beta}$) about its extent. She expects her future self to complete the number of tasks influenced by the parameter $\widehat{\beta}$ rather than β:

$$\widehat{t}(\gamma) = \arg\max_t \ wt - \frac{1}{\widehat{\beta}}C(t,\gamma). \tag{1.3}$$

For simplicity, we focus on the case when there are just two different complexity levels. In particular, the tasks can be either all Short, or all Long, which corresponds to $\gamma = \gamma_S$ and $\gamma = \gamma_L$, respectively. Therefore, when the equation 1.1 is applied to the search of the optimal number of Short tasks, it gives $t^\star(\gamma_S)$, and when the equation 1.1 is applied to the search of the optimal number of Long tasks, it gives $t^\star(\gamma_L)$. Likewise, in this case the equation 1.2, gives $t_2^\star(\gamma_S)$ and $t_2^\star(\gamma_L)$. Finally, equation 1.3 gives $\widehat{t}(\gamma_S)$ and $\widehat{t}(\gamma_L)$.

To simplify notation and get rid of the brackets, we use the following notation:

- Optimal number of tasks, from the point of view of the first period: $t_S^\star \coloneqq t^\star(\gamma_S)$ and $t_L^\star \coloneqq t^\star(\gamma_L)$.
- Optimal number of tasks, from the point of view of the second period: $t_{2S}^\star \coloneqq t_2^\star(\gamma_S)$ and $t_{2L}^\star \coloneqq t_2^\star(\gamma_L)$

- From the first period's perspective, what she believes will be the optimal number in the second period: $\widehat{t_S} \coloneqq \widehat{t}(\gamma_S)$ and $\widehat{t_L} \coloneqq \widehat{t}(\gamma_L)$.

3.2 Commitment contracts

A penalty-based commitment contract is available to the agent in period 1. It consists of a ***threshold*** denoted by #, which is a number of tasks, and a ***penalty*** denoted by \$, which is a monetary amount. The agent is offered only one commitment contract. This contract can be of one of three *types*:[5]

- **Contract for Short** applies only if the agent receives Short tasks. Suppose the agent is offered a contract for Short under piece-rate wage $w > 0$ in period 1. She can choose *not* to take up the contract. If she does so and then does $t \geq 0$ Short or Long tasks in period 2, she gets wt. Alternatively, in period 1, the agent can choose *any* positive number of tasks to be the commitment threshold $\#_S$, and *any* positive penalty to be the commitment penalty $\$_S$. We will then address this contract as $\mathbf{(\#_S, \$_S)}$. In period 2, if the agent gets Short tasks (with chance p) and does $t \geq 0$ of them, she gets $wt - \$_S \mathbb{1}\{t < \#_S\}$. If, in period 2, the agent gets Long tasks (with chance $1-p$), and then does $t \geq 0$ tasks, she gets wt.

- **Contract for Long** applies only if the agent receives Long tasks. It works analogously to a contract for Short. Suppose the agent is offered a contract for Long under piece-rate wage $w > 0$ in period 1. She can choose *not* to take up the contract. If she does so and then does $t \geq 0$ Short or Long tasks in period 2, she gets wt. Alternatively, in period 1, the agent can choose *any* positive number of tasks to be the commitment threshold $\#_L$, and *any* positive penalty to be the commitment penalty $\$_L$. We will then address this contract as $\mathbf{(\#_L, \$_L)}$. In period 2, if the agent gets Short tasks (with chance p) and does $t \geq 0$ of them, she gets wt. If, in period 2, the agent gets Long tasks (with chance $1-p$), and then does $t \geq 0$ tasks, she gets $wt - \$_L \mathbb{1}\{t < \#_L\}$.

- **Contract for Both** applies regardless of the task length. If she is offered a contract for Both under piece-rate wage $w > 0$ in period 1, she can choose *not* to take up the contract. Then, in period 2, she can do $t \geq 0$ Short or Long tasks

[5] The agent does not choose *between* contract types. Only one contract is suggested to her, and she makes her choice upon observing its type.

and get wt. Alternatively, in period 1, she can choose *any* positive number of tasks to be the commitment threshold $\#_B$, and *any* positive penalty to be the commitment penalty $\$_B$. We will then address this contract as $(\#_B, \$_B)$. In period 2, *no matter the task assignment*, she gets $wt - \$_B \mathbb{1}\{t < \#_B\}$, where $t \geq 0$ is the number of Short or Long tasks she does.

Contract for Short, Contract for Long

Either a contract for Short, or a contract for Long affect the agent only under the specified task length. Therefore, we assume that the agent's decision is based on certain costs. This assumption relies on the difference *between* Short and Long tasks in our experiment; we hope this difference will make the uncertainty *within* each length negligible. The agent' optimal commitment decision is $(\#^\star(\gamma), \$^\star(\gamma))$, such that

$$\#^\star(\gamma) = \arg\max_t \; wt - C(t, \gamma) \tag{1.4}$$

$$-\frac{1}{\hat{\beta}} C(\#^\star(\gamma), \gamma) + w\#^\star(\gamma) \geq -\$^\star(\gamma) - \frac{1}{\hat{\beta}} C(\widehat{t}(\gamma), \gamma) + w\widehat{t}(\gamma), \tag{1.5}$$

where $\widehat{t}(\gamma)$ is defined in 1.3. Notice that, under certain costs, $\#^\star(\gamma) = t^\star(\gamma)$, and taking up commitment is strictly preferable unless $t^\star(\gamma) = \widehat{t}(\gamma)$.

Similarly to the previous case, we simplify the notation by omitting the brackets. So, for Short tasks, $\#_S^\star \coloneqq \#^\star(\gamma_S)$ and $\$_S^\star \coloneqq \$^\star(\gamma_S)$. Similarly, for Long tasks, $\#_L^\star \coloneqq \#^\star(\gamma_L)$ and $\$_L^\star \coloneqq \$^\star(\gamma_L)$.

Contract for Both

Under $\widehat{\beta} < 1$, choosing $\#_B = \#_L$ and $P_B = P_L$ in the Commitment for Both is weakly better than not committing at all. By Lemma 1 in the appendix (page 41), the individual will at least $\#_B$, which gives them at least as high expected consumption as under no commitment. Thus, taking up a contract for Both is weakly better than not committing at all: the individual can either choose $\#_B = \#_L$ and $P_B = P_L$, or an at least as a good contract. Therefore, we will not discuss the decision *whether* to take up a contract for Both, but the decision of *which* contract for Both is the best to take up.

We begin by defining the contracts that ensure the agent reaches the threshold irrespective of the task length. By Lemma A.1 in the appendix, it is sufficient

to provide incentives strong enough for completing Long tasks, as these will also suffice for completing Short tasks.

Definition 1. A contract for Both $(\#_B, \$_B)$ is ***full***, if $\#_B \geq \widehat{t}_L$, and

$$-\frac{1}{\hat{\beta}}C(\#_B, \gamma_L) + w\#_B \geq -\$_B - \frac{1}{\hat{\beta}}C(\widehat{t}_L, \gamma_L) + w\widehat{t}_L. \tag{1.6}$$

We will also define the full contract that the agent prefers to any other full contract (but not necessarily to any contract in general).

Definition 1'. A contract for Both $(\#_B^\star, \$_B^\star)$ is ***optimal full***, if

1. $(\#_B^\star, \$_B^\star)$ is a full.
2. For any full contract for Both $(\#_B, \$_B)$,

$$p\left(w\max\{\widehat{t}_S, \#_B^\star\} - C(\max\{\widehat{t}_S, \#_B^\star\}, \gamma_S)\right) + (1-p)\left(w\#_B^\star - C(\#_B^\star, \gamma_L)\right) \geq$$
$$\geq p\left(w\max\{\widehat{t}_S, \#_B\} - C(\max\{\widehat{t}_S, \#_B\}, \gamma_S)\right) + (1-p)\left(w\#_B - C(\#_B, \gamma_L)\right).$$

Notably, an optimal full contract does not necessarily incentivize the agent to perform the same number of Short and Long tasks. This is because it may require fewer tasks than the agent would complete upon receiving Short tasks without any commitment (denoted by $\widehat{t}_S$).

We now turn to contracts that do not incentivize the agent to complete all required tasks if the tasks are realized to be Long.

Definition 2. A contract for Both $(\#_B, \$_B)$ is ***non-full***, if $\#_B \geq \widehat{t}_S$, and

$$-\frac{1}{\hat{\beta}}C(\#_B, \gamma_L) + w\#_B < -\$_B - \frac{1}{\hat{\beta}}C(\widehat{t}_L, \gamma_L) + w\widehat{t}_L. \tag{1.7}$$

Definition 2'. A contract for Both $(\#_B, \$_B)$ is ***optimal non-full***, if

1. $(\#_B^\star, \$_B^\star)$ is a non-full.
2. For any non-full contract for Both $(\#_B, \$_B)$,

$$p\left(w\#_B^\star - C(\#_B^\star, \gamma_S)\right) + (1-p)\left(-\$_B^\star\right) \;\geq\; p\left(w\#_B - C(\#_B, \gamma_S)\right) + (1-p)\left(-\$_B\right).$$

By definition, a non-full contract is such that the individual plans to fail it if she gets Long tasks but plans to follow through if she gets Short tasks. She might select such a contract if the benefits from boosting her output under Short tasks are high, while reaching the same output under Long tasks is very costly. If boosting the output under Short tasks requires only a moderate penalty, the benefits from boosting exceed the penalty she would have to pay upon receiving Long tasks.

Definitions 1' and 2' allow the choice of the optimal contract for Both to be the choice between optimal full and optimal non-full contracts for Both. The Proposition below establishes how the parameters of our model affect this choice.

Proposition.
Let $\mathbf{\Delta}$ *be the difference in expected consumption between the optimal non-full and optimal full contracts: it is positive if the agent prefers the optimal non-full contract to the optimal full contract and negative otherwise. Then* $\mathbf{\Delta}$

- *increases if* γ_L *increases;*
- *increases if* γ_S *increases* $\left(\text{if } \frac{\partial^3 C}{\partial t^2 \partial \gamma} > \frac{\partial^3 C}{\partial t^3} \text{ and sufficiently small } \frac{\partial C}{\partial \gamma}\right)$.[6]

Proof. See page 42 in the appendix.

3.3 Finding the Planning Motive in the Data

Definition 3. Let us consider two corresponding contracts for Both $(\#_B, \$_B)$ and for Long $(\#_L, \$_L)$. The contract for Both is **partial** if
Case 1: $\#_B > \#_L$ *and* $\$_B \leq \$_L$, *or*
Case 2: $\#_B = \#_L$ *and* $\$_B < \$_L$.

By the definition of partial contracts, they are planned to be failed if the agent gets Long tasks. We rely on the lemma below to identify them.

Lemma.
Assume that the agent always chooses the minimum penalty that allows her to follow

[6] These are the conditions that we argue to hold in our data. We discuss the effect of other combinations in parameters in the proof.

through with commitment whenever she plans.[7] *Then any partial contract for* ***Both*** *is a non-full contract for* ***Both****.*

Proof. The assumption suggests that incentive compatibility restriction (1.5) holds as equality for the contract for Long $(\#_L, \$_L)$. The statement of the lemma follows from the definition of a partial contract.

Intuitively, we assume that the penalty $\$_L$ in the contract for Long is *just enough* for the agent to reach the threshold $\#_L$ if she gets Long tasks. It means that, under the Long tasks, a higher threshold (Case 1) or a smaller penalty (Case 2) makes her fail the contract. Notably, we do not classify any contract. We cannot distinguish between partial and full contracts if $\#_B > \#_L$ and $P_B > P_L$. Also, our model cannot explain take-up of the contracts that have threshold for Long $\#_L$ higher than threshold for Both $\#_B$.

4 EXPERIMENT DESIGN

We ran our experiment at the UCSD Economics Laboratory in May 2023. Each participant came to the location on two days, one week apart. Both Day 1 and Day 2 of the experiment involved reading instructions and answering *experimental questions*: questions about commitment and additional ones. The main difference between the days was the opportunity for the participants to engage in *paid tasks* at the end of Day 2. In this section, we will cover the instructions on Day 1 and the paid tasks, and we refer the reader to the appendix for the details about Day 2.[8]

First, we describe the paid tasks, which involved transcribing lines of Greek letters that could be Short or Long. Short and Long tasks differ by the number of letters to transcribe—Long tasks are harder—but yield the same benefit. Specifically, doing one task yields a participant the same piece-rate wage, no matter its length.[9] Thus,

[7] This assumption is common in the studies on commitment (Bai et al., 2021; John, 2020). Intuitively, if the agent believes \$10 is sufficient to enforce commitment for Long tasks, she will not set the penalty to \$15. Notably, this assumption is the opposite of the individuals choosing penalties that are too small, which is regularly observed in the data.

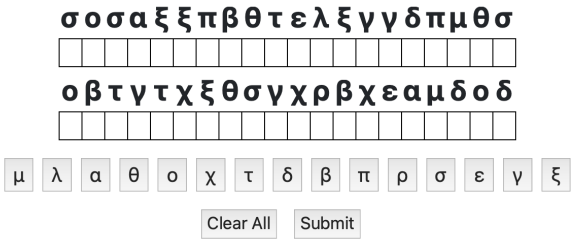
[8] The experimental questions on Day 2 are almost identical to those on Day 1. The only difference is that the tasks they concern are to be completed on the same day rather than in one week. Thus, the distance between responding to the questions and doing the tasks is one week on Day 1 and about 30 minutes on Day 2. For the results on Day 2, we also refer the reader to the appendix, as we do not discuss them in the main text.

[9] We use different approaches to payment in the additional questions. However, neither of the additional questions concerns commitment, and the participants see them only after they answer the commitment question.

Figure 1.1: Examples of Paid Tasks

(a) Example of a Short task

(b) Example of a Long task



Note: The figure gives examples of a Short task and a Long task. The participants were told to click on the keys with the corresponding images one by one. Once they have filled all the cells, they could click "Submit" to complete the task. They had an option to clear all cells by clicking "Clear All." A Short task has 15 letters, and the participants could enter at most one letter incorrectly. A Long task has 40 letters, and the participants could enter at most four letters incorrectly. Both numbers of mistakes allow the participants to achieve 90% accuracy by replacement.

uncertainty in task length exogenously introduces uncertainty in cost. Secondly, we describe the experimental questions. Some of them offer commitment contracts, and the others are additional and not related to commitment. Thirdly, we elaborate on the timeline and the payments. Finally, we describe the sample by elaborating on the number of participants, the attrition rate, and how we managed the outliers.

4.1 Paid Tasks

We distinguish between *experimental* (or paid) *tasks*, which are paid activities for the participants to engage in, and *experimental questions* (or just *questions*), which require a response. Our participants engage in transcribing Greek letters, similar to the tasks introduced by (Augenblick, Niederle, and Sprenger, 2015; Augenblick and Rabin, 2019). One difference between our study and the previous ones is that the participants could face either Short or Long tasks (see Figure 1.1). The chance of receiving either all tasks as Short or all tasks as Long creates the exogenous variation in costs. Another difference between our study and the previous ones is that we made the letters clear instead of blurry. Given the long time of the experimental sessions, we clarified the letters to prevent the participants from getting a headache or experiencing any side effects.

The participants learned at the very beginning that they would receive an opportunity to complete some amount of these tasks to receive payment on top of payment for

participation. They also engaged in practicing the tasks (we will elaborate on the practices while explaining the timeline).

4.2 Experimental Questions

The participants answered questions of four different types on both days. These questions concern their paid task experience: some of them are commitment options, and others can affect the type of tasks the participant gets or how much she would be paid. One of the questions from either of the two days was selected for them as the *question-that-counts*. This would be the only question that actually affects their paid tasks. The participants were strongly encouraged to treat each question as if it were the one that counted.

Commitment Questions

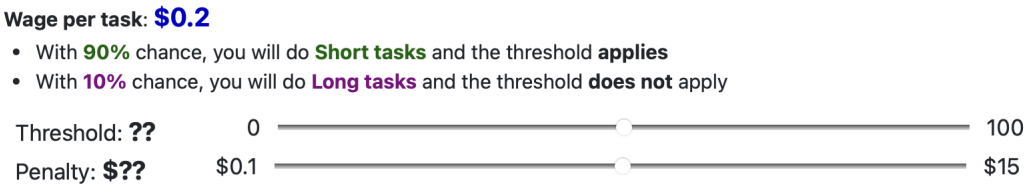
The first part of the experimental questions offered commitment options to the participants. To make the instructions simple, we explained the options in terms of *thresholds* and *penalties*. We ensured that the participants paid close attention to the instructions by reading them out loud (only until the beginning of the quiz on the commitment questions).

First, we reminded the participants that a question from this part could become the question-that-counts. In that case, they would get all of their tasks to be Short with a 90% chance and all of their tasks to be Long with a 10% chance.

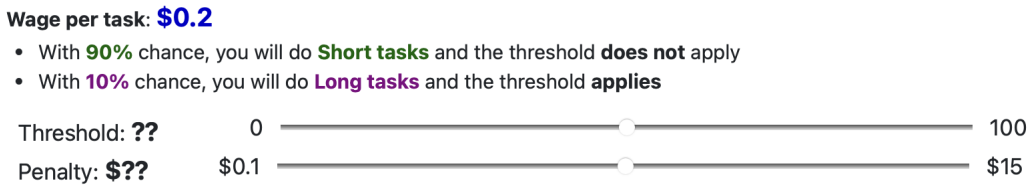
Then, we explained the concept of thresholds and penalties: the threshold is the number of paid tasks that a participant needs to complete on Day 2 to avoid losing the amount of money specified in the penalty. The questions in this part varied by two components: the wage per task that the participant would get if this question became the question-that-counts and when the threshold applies. Thresholds in some questions would apply only to Short tasks, while others would apply only to Long tasks. There were also questions about thresholds being applied no matter the length of the tasks. We made a few details explicit in our instructions, the most important being that they *do not have to* set any threshold. We mentioned it at the beginning of the instructions; our comprehension quiz contained a corresponding example, and there was a reminder on each screen with the questions. Further, we emphasized that the thresholds do not affect the number of letters in the tasks and that reaching the threshold means they lose no penalty.

Figure 1.2: Examples of Commitment Questions

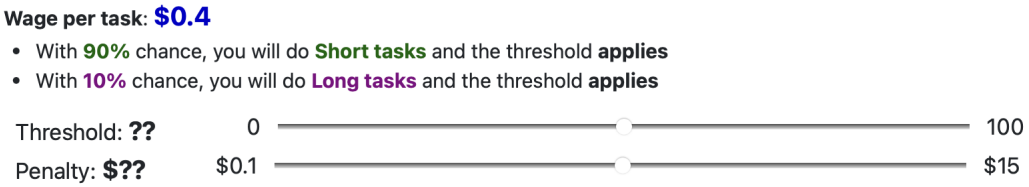
(a) Example 1



(b) Example 2



(c) Example 3



After giving the instructions, we showed the participants the examples (see Figure 1.3) to ensure that they understood that both the wage and when the threshold applies can change between the questions and that the chances of receiving Short and Long tasks are 90% and 10%, respectively, in all of them.

Additional Questions

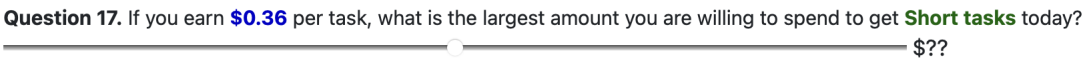
Willingness-to-Pay

We told the participants that we wanted to know how valuable receiving Short tasks on Day 2 was to them. If a question from this part becomes the question-that-counts, they would get Long tasks by default. We asked them about the largest amount they would be willing to pay on Day 2 to receive Short tasks instead (see Figure 1.3a).

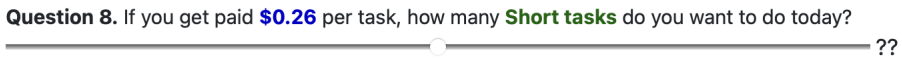
We explained that responding honestly was their best chance to receive a larger payment. The incentives for these questions were as follows. First, we would randomly set a price from \$0 to \$30 for the switch from Long tasks, which they have by default, to the Short ones. If this price is equal to or smaller than the response,

Figure 1.3: Examples of Additional Questions

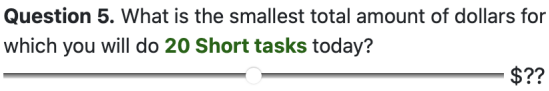
(a) Willingness-to-pay



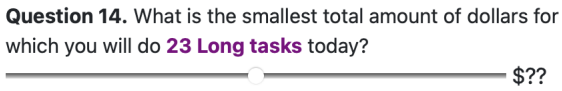
(b) Task decisions



(c) Payment decisions: Short tasks



(d) Payment decisions: Long tasks



they get Short tasks, and we deduct this price from their completion payment. If the price exceeds this amount, they keep Long tasks.

Task decisions

Task decisions are very similar to the questions by Augenblick and Rabin, 2019 about how many tasks the participants would want to do. We asked the participants about how many Short tasks they wanted to do on Day 2 (see Figure 1.3b). The reason why we did not ask about Long tasks is that we did not want our participants to be cautious about the number of tasks they want to do for the sake of avoiding discomfort for their future self. The participants were informed that they would not be able to do more tasks than the number they responded. If they would do fewer tasks, they would receive only $20 for the whole experiment instead of $50 completion payment plus the money for the paid tasks.

Payment decisions

We asked the participants about the minimum amount they would have to be paid to do the tasks. Each question would give a number of Short tasks (or a number of Long tasks), and suggest choosing the smallest TOTAL amount of dollars the participant would accept for performing these tasks (see Figures 1.3c and 1.3d).

If a question from this part becomes the question that counts, the payment for doing the required amount of tasks was determined as follows. We would randomly choose the amount of dollars we pay the participant for completing the tasks. Then, if this

amount is less than the response, the participant would not be allowed to do any tasks (she would only receive her completion payment). If the randomly selected amount is larger than the response, it is this amount the participant receives for doing the tasks. Similarly to incentivizing the task decisions, we would pay the participants only $20 if they fail to do the required amount of tasks.

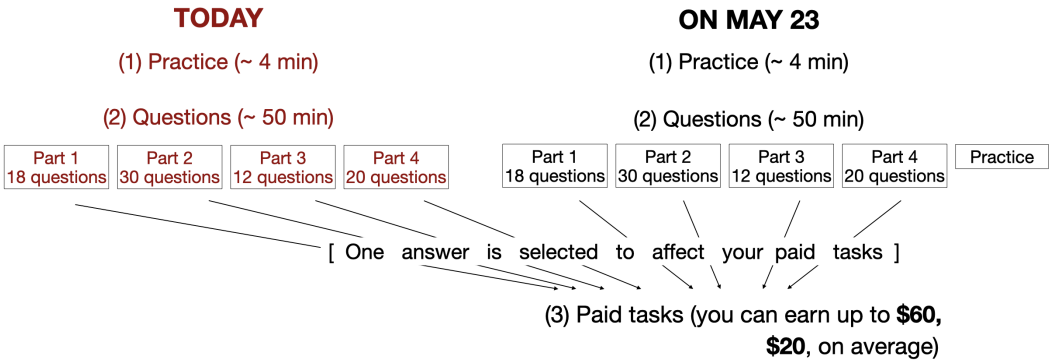
4.3 Timeline and Payments

The participants came to the location on two days, one week apart. We call these days Day 1 and Day 2. At the beginning of Day 1, the participants were required to confirm their eligibility for the study. The eligibility criteria were a part of the study announcement. The participants needed to confirm that they would participate on Day 1 and Day 2 and would be willing to receive all of the payment in cash after the session on Day 2. We give the full list of eligibility criteria in the appendix. Further, the participants were suggested to agree to the informed consent terms to proceed with the experiment (see the appendix).

After the participants had agreed to the informed consent terms, they proceeded to the instructions. As we had told them upon their entrance to the lab, we read the instructions out loud (until the beginning of the experimental questions). We first explained the concept of *paid tasks*—the tasks that they had an opportunity to complete for the payment on top of $50 for completing both days of the study and following through with their decisions.

Then, we showed the participants the timeline of the experiment. Figure 1.4 presents the exact timeline that we showed the participants on May 16, which was Day 1 for them. When we showed them the timeline, we first suggested they look through the activities on May 16 (in red, in the left part of the timeline). Day 1 began with the Practice. Firstly, we reviewed the paid tasks: number of letters, examples, etc. Secondly, the participants were required to complete one Short task and one Long task. Just as in the paid tasks, they were allowed to make up to one mistake in a Short task and up to four mistakes in a Long task. After the practice, the participants were to answer the experimental questions. After they answered the questions on Day 1, they were free to leave the lab. On Day 1, we described Day 2 only briefly by mentioning that they would engage in similar activities and then have an opportunity to engage in paid tasks. We told them we would elaborate on how their answers to our questions would affect their experience with paid tasks after the Practice.

Figure 1.4: Experiment Timeline (presented on May 16)



After the participants completed the Practice, we told them that one of the experimental questions (from either of the two days) would be selected as the *question-that-counts*. We explained that their responses could not affect which question becomes the question-that-counts, and every question had an equal chance to be selected as the question-that-counts. We assured the participants that we would inform them on which question is the question-that-counts just before they begin the paid tasks.

Each part of the experimental questions began with instructions. As the commitment questions were the least straightforward, we kept reading the instructions aloud to ensure the participants spent enough time comprehending them. The instructions about commitment questions were followed by examples (Figure 1.2). Before proceeding to the commitment questions, the participants had to pass a comprehension quiz with an unlimited number of attempts. At this point, we stopped reading the instructions aloud and suggested the participants proceed independently. After the quiz, we ensured that the participants understood the correct answers by providing a screen with explanations. The Willingness-to-pay questions also included a comprehension quiz and explanations.

After the participants returned to the lab on Day 2, we essentially repeated Day 1 for them, from reading the instructions aloud to asking them to respond to comprehension quizzes. We did that to ensure that their experience across the two days was similar. After the participants had responded to all experimental questions on Day 2, they were told which question had been selected as the question-that-counts. Then, they could engage in the paid tasks to the extent the question-that-counts allowed (we described the content of the questions above). For example, suppose a question

from Part 4 (Payment decisions) was selected as the question-that-counts for a participant. Then, the content of the question determines whether the participant gets Short tasks or Long tasks. Further, a random payment from $0 to $60 was selected. If it was smaller than the participant's response, the participant was not allowed to engage in paid tasks and could only receive her $50 completion payment. If it was larger than the participant's response, she was allowed to do the amount of tasks specified in the question. If she failed to do so, she would receive $20 in total. If she completed the required amount of paid tasks, she would get the $50 completion payment and the randomly selected payment.

When the participants concluded their participation on Day 2, they could leave the lab independently. They were given the amount of money they had earned in cash at the lab exit.

4.4 Sample

We conducted our experiment at the UCSD Economics Laboratory in May 2023. We began the experiment more than a month into the spring academic term and ended it two weeks before the final exams. All our participants are at least 18 years old and are verified members of the subject pool. We recruited four groups of participants. On May 13, 44 participants joined the session to continue on May 20. On May 16, 36 more participants had their Day 1 to return on May 23.[10] Most participants returned for Day 2: 43 on May 20 and 33 on May 23. All of them apart from one[11] completed the experiment in full and received their $50 completion payment. The distribution of their bonus payments had the average of $22, median of $16, and maximum of $60.

We randomized the order of the experimental questions randomly assigning each participants to one of four groups. Table 1.1 describes the randomization. For commitment questions, we varied the order of types of contracts on one screen and the order of piece-rate wages between the screens. For the willingness-to-pay questions and task decisions, we randomized if the participants saw the piece-rate wages in increasing or decreasing order. For payment decisions, the number of tasks to perform was also either in increasing or decreasing order.

[10] One more participant on May 16 was asked to leave about 20 minutes after the beginning of the experiment as she was distracted by her smartphone.

[11] She was randomly assigned to do 32 Long tasks for $38, while the minimum payment she had been willing to accept was $35.

Table 1.1: Randomization of experimental questions

	Commitment contracts (order on screen)	Commitment piece-rate wage (between screens, ¢)	Additional questions (piece-rate wage/number of tasks)
Group 1	Short, Long, Both	20, 30, 40, 60, 10, 50	Increasing
Group 2	Both, Long, Short	20, 30, 40, 60, 10, 50	Increasing
Group 3	Both, Short, Long	50, 10, 60, 40, 30, 20	Decreasing
Group 4	Long, Short, Both	50, 10, 60, 40, 30, 20	Decreasing

Our main subject pool consists of 69 participants. As we had pre-registered,[12] we excluded the participants with little to no variability in their responses to non-commitment questions. The reason for that is the large variability in the piece-rate wages in Parts 2 and 3 (from \$0.1 to \$0.54 or more), or in the number of tasks to do in Part 4 (from 5 to 32). We excluded those who did not vary their responses within two out of three parts. We also pre-registered a criterion for excluding the participants who failed three or more attempts on either of the comprehension quizzes. We dropped this criterion as we underestimated the quiz's difficulty for the participants: 12 out of 69 participants in our main sample required 3 or 4 attempts. We kept them in the main sample as we clarified the correct answers to the quizzes after the participants passed them. Another reason for many mistakes could be that we emphasized that the participants had an unlimited number of attempts.

5 RESULTS

We present our results in the following order. First, we briefly describe commitment take-up, as shown in Figure 1.5. Second, we identify the proportion of partial contracts (those planned to fail) in the data. Third, we provide evidence that our partial commitment measure accurately captures contracts that are planned to fail.

Notably, in this section, we will only discuss the contracts *designed* by the participants, rather than generic commitment contracts. Thus, a contract for Short $(\#_S, \$_S)$ is the contract that was selected by a participant among other alternatives she had. In Section 3, we used this notation for a generic contract for Short, not necessarily selected by the participant. Similarly, we will denote taken-up contract for Long by $(\#_L, \$_L)$, and taken-up contract for Both by $(\#_B, \$_B)$.

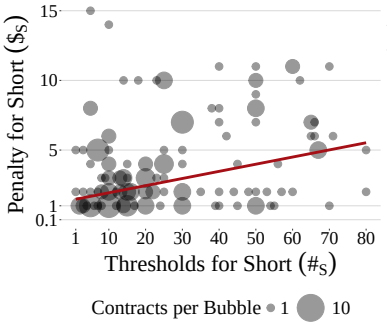
[12] We pre-registered our study on the AsPredicted platform, https://aspredicted.org/J76_BFH.

5.1 Commitment Take-up

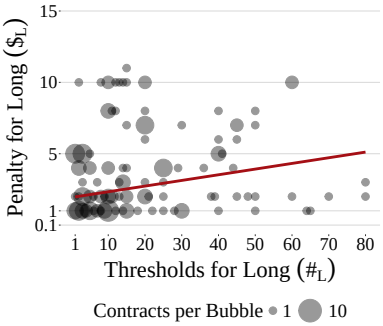
Figure 1.5 demonstrates the commitment take-up by contract type: for Short, for Long, and for Both. In total, the participants were suggested 414 of each type. They took up 254 (61.3%) contracts for Short, 234 (56.5%) contracts for Long, and 250 (60.3%) contracts for Both. We give more details on commitment take-up in Appendix A.4

Figure 1.5: Commitment Take-up by Contract Type

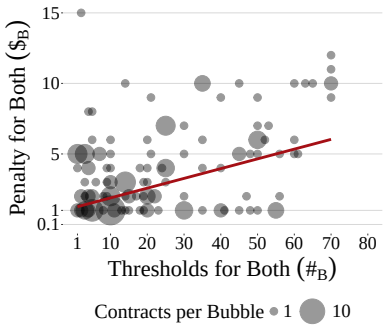
(a) Contracts for Short



(b) Contracts for Long



(c) Contracts for Both



Note: Each panel displays the relationship between penalty and threshold levels for a different type of commitment contract: (a) contracts that apply only to Short tasks, (b) contracts that apply only to Long tasks, and (c) contracts that apply to both task types. In each plot, one bubble represents a group of similar contracts; the bubble size reflects the number of such contracts (with a maximum of 10 per bubble). The red line indicates the linear correlation between penalty size and threshold level.

5.2 Partial Contracts

Unless stated otherwise, all three contracts are considered under the same piece-rate wage.[13] When two contracts of any type share the same piece-rate wage, we refer to them as *corresponding* contracts (or we say they *correspond*). Additionally, when discussing a contract for Both, we use the notations $\#_S > 0$ (or $\#_L > 0$) to indicate that the individual is also committing to the corresponding contract for Short (or Long).

Consider a contract for Both $(\#_B, \$_B)$ such that $\#_L > 0$ (the participant takes up the contract for Long for the corresponding wage). Based on Lemma on page 16, we identify this contract as *partial*, if

$$\#_B - \#_L > 0 \wedge \$_B - \$_L \leq 0, \text{or} \tag{1.8}$$

$$\#_B = \#_L \wedge \$_B - \$_L < 0. \tag{1.9}$$

We define partial contracts on the corresponding subsamples of total commitment take up. For example, no contract for Both with $\#_B = 1$ can be partial as $\#_L \geq 1$, which makes us exclude such contracts from consideration. In total, 69 participants took up 250 contracts for Both (out of 414). We identify partial contracts on the following sample of contracts for Both:

$$\{(\#_B, \$_B)\colon \#_B > 1, \#_L > 0\},\ n = 200.$$

As Figure 1.6 shows, the proportion of partial contracts is **38%**, which is statistically different from zero.[14] The proportions remain stable across different piece-rate wages: it is between 33 and 43% across the six piece-rate wages we consider. Further, restricting the requirements does not change the proportion of partial contracts extremely. For example, considering contracts $\{(\#_B, \$_B) : \#_B - \#_L \geq 16 \wedge \$_B - \$_L \leq 0\}$ on the sample $\{(\#_B, \$_B)\colon \#_B > 16, \#_L > 0\}$ suggests the proportion of 29% (with [13, 39.5] 95% confidence interval).

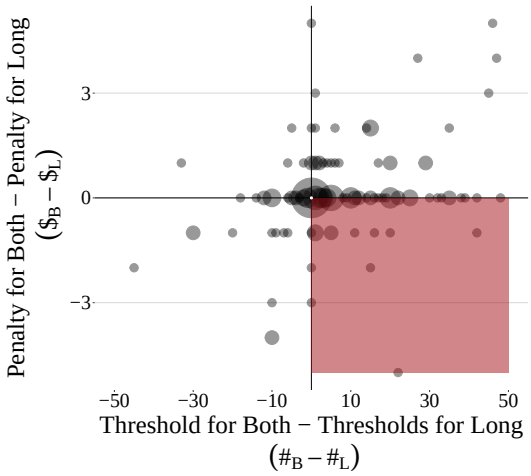
Figure 1.6 demonstrates that a large portion of the results is driven by the contracts for Both that have the same penalty as the contracts for Long, but a higher threshold.

[13] The piece-rate wage w refers to the monetary reward a participant receives for completing one task. This wage remains constant regardless of task length or whether the participant commits to the task.

[14] Arguably, we should consider 25% as the benchmark for the proportion of partial contracts due to the symmetry of the identification requirements across the two axes. The proportion of 38 percent is different from 25 at 5% significance level.

Figure 1.6: Partial Contracts, %: **38 [27.4, 48.6]**

$$\begin{bmatrix} \#_B - \#_L > 0 \wedge \$_B - \$_L \leq 0, & \text{or} \\ \#_B = \#_L \wedge \$_B - \$_L < 0. \end{bmatrix}$$



Note: The figure displays the commitment take-up, where red-shaded areas delineate the regions associated with partial contracts. The subtitles give the proportion of contracts for Both, which are identified as partial under the corresponding definition, along with a 95% CI with standard errors clustered by individuals. One observation is a contract for Both; a bubble represents a group of similar contracts for Both. A larger bubble corresponds to a larger number of similar contracts. The x-axis denotes the difference between the threshold in the contract for Both and the threshold in the contract for Long under the same piece-rate wage ($\#_B - \#_L$). The y-axis denotes the difference between the penalty in the contract for Both and the penalty in the contract for Long under the same piece-rate wage ($\$_B - \$_L$). The details of the partial contract definitions are provided below the subtitles. The partial contracts are identified among the 200 contracts for Both that have thresholds above 1. We omit 11 out of 200 contracts from the figure as outliers.

These contracts are located on the x-axis in the right part of the Figure. It is important to note that if participants are not sensitive to penalties, we cannot be certain that such contracts are planned to fail.[15] However, we have evidence that participants are sensitive to penalties. They often select penalties larger than \$1, and, more importantly, the thresholds they choose are positively correlated with the penalties. Therefore, the equality of penalties does not raise a concern, especially given the large difference in thresholds.

[15]This idea can be illustrated by an extreme case in which individuals *always* follow through whenever they commit. In this case, despite the difference in thresholds, the partial contracts are not planned to fail.

5.3 Partial Contracts Capture Planning-to-Fail

We now want to show that the participants selected partial contracts intentionally rather than by chance. In this subsection, we elaborate on the following three arguments.

1. Random contract selection cannot account for the full amount of partial contracts.
2. Strengthening the threshold requirements in the definition of partial contracts has only a moderate effect on their proportion.
3. Proportion of partial contracts is positively correlated with the costs of doing tasks. Notably, the proportion of contracts that are planned to be failed upon receiving Long tasks is positively correlated with the costs based on Proposition on page 16.

Partial contracts do stem from random choices of contracts

First, if selection of thresholds and penalties was completely random, comparison of contracts for Both and for Short would give the same results and the comparison of the contracts for Both and for Long, which we do in Figure 1.6. Figure 1.7a repeats Figure 1.6 so it is easier to compare it to Figure 1.7b.

Threshold requirements only moderately affect partial contracts

We want to demonstrate that restricting the requirement for the difference between contracts for Both and for Long in thresholds. Currently, we consider the difference in 1 task to be sufficient, if the penalties are the same (or lower in the contract for Both). We will consider partial-$\boldsymbol{X}$ contracts for Both, defined as follows.

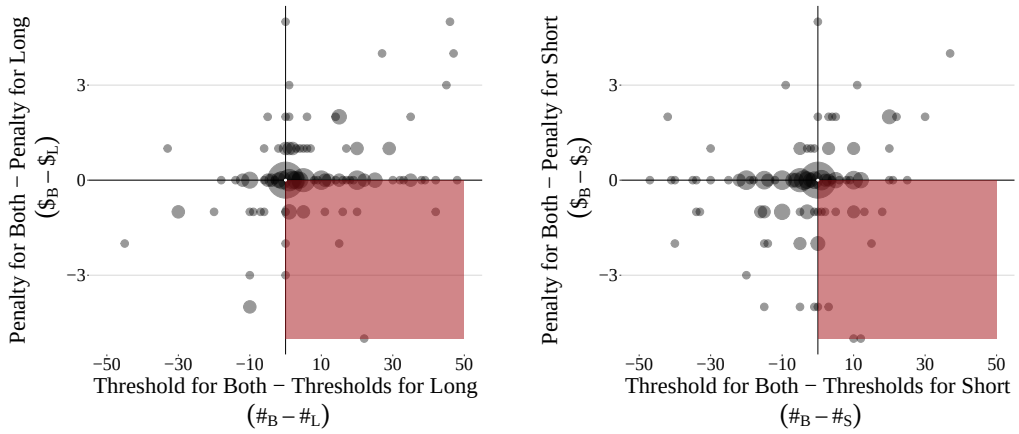
Definition 3. A contract for Both is partial-$\boldsymbol{X}$, $\boldsymbol{X} > 0$ if

- $\#_B > X$,
- $\#_B - \#_L \geq \boldsymbol{X}$ and $\$_B - \$_L \leq 0$.

Figure 1.7: Both vs. Long and Both vs. Short: All Contracts for Both

(a) Partial Contracts, %: **38 [27.4, 48.6]**
$$\begin{bmatrix} \#_B - \#_L > 0 \wedge \$_B - \$_L \leq 0, & \text{or} \\ \#_B = \#_L \wedge \$_B - \$_L < 0. \end{bmatrix}$$

(b) 'Partial' Contracts, %: **20.5 [12.5, 28.6]**
$$\begin{bmatrix} \#_B - \#_S > 0 \wedge \$_B - \$_S \leq 0, & \text{or} \\ \#_B = \#_S \wedge \$_B - \$_S < 0. \end{bmatrix}$$



Note: This figure illustrates the argument that the proportion of partial contracts cannot be attributed to noisy responses. The left panel replicates Figure 1.6—with partial contracts in the red area—and should be compared to the right panel. In the right panel, we repeat the estimation of the share of partial contracts, but use contracts for Short instead of contracts for Long. If participants selected contracts in a noisy manner, contracts for Long and contracts for Short would exhibit similar patterns. Therefore, we would expect to observe the same proportion of contracts in the red area in the right panel as in the left one. However, the proportion of contracts that are classified as partial is significantly lower, indicating that the observed pattern is unlikely to be driven by noise.

Thus, $\boldsymbol{X}$ is the minimum difference $\#_B - \#_L$ required for a contract to be partial-$\boldsymbol{X}$. The definition suggests that any partial-$\boldsymbol{X}$ contract is also partial. Notably, partial-$\boldsymbol{X}$ contracts are properly defined only on a subsample of contracts for Both:

$$\{(\#_B, \$_B) \colon \#_B > \boldsymbol{X}, \#_L > 0\}.$$

The proportion of partial-$\boldsymbol{X}$ contracts in the corresponding sample is smaller than the proportion of partial contracts 38% (see Figure 1.6). However, reasonable values of $\boldsymbol{X}$ do not make the proportion of contracts we document as non-full smaller than 30%. In particular, the minimum $\boldsymbol{X}$ required is 16 tasks: there are **29%** [13, 39.5][16] partial-$\boldsymbol{X}$ contracts among 99 (such that $\#_B > 16$ and $\#_L > 0$).

[16] We find the 95% confidence interval under individual fixed effects and report it in parentheses.

Partial contracts have the same features as non-full contracts

We defined partial contracts and, by definition, they *look like* the contracts that are planned to be failed upon getting Long tasks (non-full contracts). We want to ensure that they indeed capture the planning-to-fail motive. The Proposition on page 16 suggests that the proportion of the contracts planned to be failed should be positively correlated with the costs of doing the tasks. Notably, we can expect the costs of doing Short and Long tasks to be positively correlated. Thus, we rely on the following variables to be correlated with the proportion of non-full contracts ([.] gives the correlation sign):

1. [−] How many Short tasks do you want to do?
2. [+] How much do you want to be paid for doing X Short tasks?
3. [+] How much do you want to be paid for doing X Long tasks?

We believe all three measures to be exogenous, as we collected them independently from commitment decisions in a very different environment. They were also separated by blocks of instructions and one comprehension quiz. If our definition of partial contracts captures at least a part of the non-full contracts, their proportion should correlate with the measures above.

As we proceed, we pool the data that we collected one week in advance with the data we collected just before the participants could engage in the paid tasks. We asked them the same questions as one week in advance, as the timeline in Figure 1.4 suggests. The proportion of partial contracts taken-up on the same day was **29.75%** $<$ 38.0% (p-value 0.1 without clustering by ID). We attribute the difference to the noisy take-up one week in advance (what Kaur et al., 2015, call 'costly experimentation').

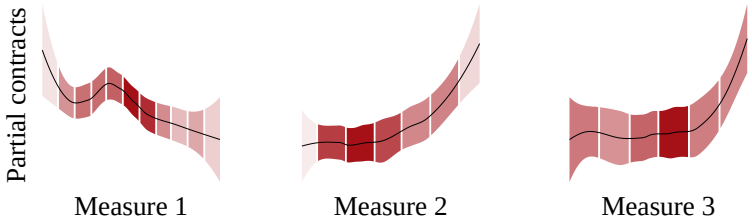
Although we pool the contracts from the two days together, we want to be transparent that the take-up right before paid tasks drives most of the results we report below. We attribute that to the noisy take-up one week in advance: considering more restrictive definitions (in particular, more restrictive with respect to the difference in thresholds between the contracts for Long and for Both) makes the result one week in advance align with that for take-up right before the paid tasks.

We conduct the correlation analysis on the contract level by defining the dummy for a partial contract and correlating it with the corresponding measures after controlling

for piece-rate wage w of a contract and the day it was picked up. We cluster standard errors by individuals. Table 1.2 reports the results of the correlation analysis and illustrates the significant correlations.

Table 1.2: Correlations between Partial Contracts and Measures of Costs

Partial Definition	ℂorr(., Measure 1)	ℂorr(., Measure 2)	ℂorr(., Measure 3)
Standard	(a) $-^{**}$	0	0
Higher # difference	0	(b) $+^{**}$	(c) $+^{**}$
Higher $ difference	0	0	0



- The figures illustrate the significant relationships on the contract level.
- Specifically, they show non-parametric analysis of residuals (we skip the axes ticks).
- Higher color intensity corresponds to more observations in the area.

Overall, the relationships that should be held between the partial contracts and the measures of costs (negative for Measure 1, positive for Measures 2 and 3) are never rejected at 0.1 significance level, and sometimes supported at 0.05 significance level. Therefore, we can rely on our measure to capture the planning-to-fail motive.

6 DISCUSSION

The results of our experiment suggest that an exogenous intervention into costs makes the individuals adapt their commitment contracts accordingly. Our estimates of the proportion of contracts that participants plan to sometimes fail vary from around 30% for the sample in general, and up to 50% for the subsample of individuals that strongly distinguish between the Short and Long tasks. Notably, there is a large portion of contracts that we cannot classify, and therefore, these proportions are likely to be even higher. In this section, we discuss the impact of our results and details of the analysis.

The vast scale of anticipated failures in our sample suggests that the participants could have planned many of the commitment failures of the previous studies. The crucial difference between planned and unplanned failures suggests exploring if planned failures could be dominating. In the Appendix, we build a simple model of a penalty-based commitment. In a nutshell, the chance that a committed individual attributes to failure cannot be larger than the chance she believes commitment to make a difference (make her do the action while otherwise, she would not). As we can approximate the chance commitment contracts make a difference in the previous studies, we can explore the upper bound on the share of the contracts that were failed due to planning.

Table 1.3 constructs the upper bound for planned failure based for ten previous studies. Columns 1 and 2 name the study and briefly describe the setup. Column 3 approximates the chance the participants who committed would have done the required action without commitment. We assume that the individuals who take up commitment have the same chance of fulfilling the goal without commitment as those who are not offered commitment at all. Therefore, we use the proportion of those doing the action among the individuals who were not offered commitment at all. We rely on this assumption, as it is the closest approximation available. Also, it is not clear if the frequency in the control group underestimates or overestimates the baseline for those who take up commitment. Overestimation can stem from time-consistent and naive people who do not expect to ever fail, and therefore are indifferent between taking up commitment and not. Underestimation can come from people who do not normally do the action, as it is them who require additional incentives to do so.

Column 4 of Table 1.3 gives the proportion of committed participants who followed through with the commitment requirements. The difference between this proportion and the base rate in Column 3 approximates the maximum frequency of failures that can be explained by planning. We report it in Column 5. Naturally, the planned failures rate cannot exceed 50%. Further, we can compare this rate with the overall failure rate, which we report in Column 6.

Comparison between Columns 5 and 6 shows that in 4 out of 14 treatments, planned failure might explain all or almost all failures, and it can potentially explain at least a third in 5 more.[17] It makes exploring planned failure compelling as this comparison

[17] In addition, two treatments in the study by Exley and Naecker, 2017 are consistent with the idea of planned failure as due to the suggested explanation: "... the student leaders for whom these

could have revealed no scope for its explanatory power. Therefore, although we do not claim that planned failure is necessarily as extensive as the model predicts, we emphasize that it can be a strong explanation of failures.

We now want to discuss the assumptions we make in our analysis. By the Lemma on page 16, our identification strategy relies on the assumption about the contracts for Long. We assume that, if a participant takes up such a contract, the penalty she chooses is just enough to enforce reaching the threshold under the Long tasks. Based on that, we suggest that a contract for Both with a larger threshold and the same penalty (or a smaller penalty and the same threshold) is supposed to be failed upon assignment of Long tasks. Ultimately, this assumption does not affect our results much because our identification relies mainly on the thresholds of the contracts.

There are two reasons for that. First, as Figure 1.6 suggests, most differences in penalties between contracts for Long and contracts for Short lie within \$1. The smallness of the amount and opportunity to set the penalty (using a slider) up to \$15 allows us to attribute most of these differences to mistakes. Second, we extensively discuss the approaches that require a considerable difference between thresholds for Both and for Long for a contract to be considered partial. This difference can compensate for most discrepancies in penalties between the contracts.

Our theoretical framework relies on the assumption that the uncertainty of costs *within* Short and Long tasks is negligible compared to the discrepancy *between* the two lengths. Suppose a participant takes up commitment for Long and the costs of doing Long tasks are uncertain. Still, there should be a cost realization for which penalty P_L in commitment for Long is the minimum penalty that ensures reaching threshold $\overline{e}_L$. Then a partial contract will make the individual fail for these cost realizations. Therefore, our identification approach is still indicative of an intention to fail with a positive chance. The uncertainty of costs within each category prevents us from estimating this chance, which would otherwise be 10%. On the other hand, the discrepancy between the thresholds and the task decisions highlights the potential impact of cost uncertainty on commitment decisions. Our commitment contracts are flexible enough to accommodate low thresholds, but it remains unclear whether contracts with higher thresholds would be in demand and to what extent planned failure occurs within each task length.

---

workshops are intended, are well aware of their overbooked schedules." Abandoning a part of \$15 might indicate that they did not want to receive the money for free rather than use it as a commitment device.

Table 1.3: Comparison of planned failure capacity with observed failure rates

(1) Study	(2) Setting, Number of Committed	(3) **BASE**, %	(4) **COMPLETE**, %	(5) **PLANNED CAP**, %	(6) **FAILURES**, %
Bai et al., 2021	Prepayment for doctor visits (lose money if miss, define follow-through by at least one visit)				
	Fixed commitment, $n = 39$	4	38	34	62
	Personalized commitment, $n = 40$	4	30	26	70
	Fixed commitment, discount, $n = 72$	8	32	24	68
	Personalized commitment, discount, $n = 112$	8	23	15	77
Burger and Lynham, 2010	Bets on weight loss, $n = 51$	–	20	20	80
Carrera et al., 2022	Loss of \$80 unless go to the gym 12 or more times, $n = 556$	22	65	43	35
Exley and Naecker, 2017	Loss of \$X (up to \$15) of Amazon gift card unless attend a workshop				
	Private \$X, $n = 29$	55	52	0	48
	Public \$X, $n = 31$	55	58	3	42
Giné et al., 2010	Loss of deposit if keep smoking, $n = 83$	9	34	25	66
John, 2020	Installment-savings account with self-chosen penalty, $n = 114$	–	45	45	55
Kaur et al., 2015	Self-chosen target for dominated wage contract ($n = 8,240$ worker-days)	92	97	5	3
Royer et al., 2015	Loss of stakes if not go to gym 14 days in a row, $n = 43$	20+	63	43-	37
Schilbach, 2019	Loss of money upon drinking alcohol, $n \approx 2,000$ person-days	39	53	14	47
Schwartz et al., 2014	Loss of cash-back if buy too little healthy food, $n = 632$	23	34	13	66
Average		24.2	46	22.1	54

Note: Column 3 (**BASE**) gives the chance of completing the commitment requirement when commitment is not offered (ϕ in expression (1.21)). We collect this chance from the control groups or their analogs (see the main text for the assumption we make). Column 4 (**COMPLETE**) gives the rate of follow-through on commitment ($\phi + \xi$ in expression (1.21)). Column 5 (**PLANNED CAP**) approximates the share of participants who meet the goal but would not do so without commitment (ξ in expression (1.21)). By Corollary 1, it serves as an upper bar on the chance of planned failure. Column 6 (**FAILURES**) gives the rate of failures observed in the study. We assess the explanatory power of planned failure by comparing Columns 5 and 6. Bai et al., 2021: The authors state the range of failures to be between 62% and 77%. We cite the follow-throughs by treatment after John, 2020. Royer et al., 2015: Just over 20% of participants without commitment (but with incentives) attended the gym at least once per week; the authors do not give the information about once every two weeks. Schilbach, 2019: The author does not report any specific follow-through rate on commitment. We derive it from Table 3 of the paper, and the calculations are consistent with those by Derksen et al., 2021. John, 2020 reports 37% of failures instead of 47%, which would be more in favor of planned failure explaining them. John, 2020: there is no information on defaults in the control group, but the average savings of 27 pesos compared to 429 under commitment suggests that the baseline can be mostly ignored. Despite that Bhattacharya et al., 2015 study penalty-based commitment contracts, we do not cover them as they do not report the failure rates.

7 CONCLUSION

Commitment contracts are designed to help individuals who are aware of their time inconsistency overcome their inclination to prioritize immediate gratification over long-term goals. Our study focuses on failures of penalty-based commitment contracts, which impose monetary penalties if individuals fail to perform a specified action, such as exercising or saving money. We examine whether individuals might take up commitment contracts with a plan to fail if the uncertain costs of following through turn out to be high. Motivated by potential gains from lower cost realizations, this reason for failure is underexamined compared to failures due to take-up mistakes. However, the distinction between planned and unplanned failures is crucial for the effectiveness of commitment contracts: unlike planned failures, unplanned ones can undermine the benefits of commitment contracts to the extent that they might be better off not being offered at all.

We find extensive take-up driven by the planned failure motive: from 30% to 40% of overall commitment take-up. We identify this proportion through a lab experiment in which we exogenously introduce uncertain costs by having participants transcribe Short or Long lines of Greek letters. The results of the robustness checks suggest that the proportion is not sensitive to identification details.

Although we distinguish between planned and unplanned failures, they can interact closely. Underestimating the scale of one's present bias can lead to overestimating the chance commitment changes the outcome to the desired one. Since the chance of planned failure is capped by this chance, the underestimation of present bias can be exacerbated by allowing for a larger chance of failure.

One direction for future research is estimating the scale of planned failures. Even though asking individuals about the likelihood that they will follow through with a commitment is not incentive-compatible, asking participants about the chances of others can be incentivized. Although other individuals are likely to miss individual variation in costs, using their opinions for estimation can benefit from their clearer perspective on the present bias of others (Fedyk, 2016). Further, this approach can help estimate costs without introducing exogenous variation.

The small penalties most participants selected for their commitment contracts in our study suggest another direction. On the one hand, this can be a consequence of participants considering themselves almost time-consistent. On the other hand, non-monetary consequences (such as image impact) could substitute for monetary

penalties for our participants. Regardless of the reason, we find these results promising for using commitment contracts with little to no monetary penalties. Not only can they be preferred by participants, but they also make participants lose little to no money upon failure. The study by Derksen et al., 2021 is one in this direction, and we expect more to appear in the future.

References

Acland, D., & Levy, M. R. (2015). Naiveté, Projection Bias, and Habit Formation in Gym Attendance. *Management Science*, *61*(1), 146–160. https://doi.org/10.1287/mnsc.2014.2091 (cit. on p. 8).

Afzal, U., D'Adda, G., Fafchamps, M., Quinn, S., & Said, F. (2019, May). *Implicit and Explicit Commitment in Credit and Saving Contracts: A Field Experiment* (tech. rep. No. w25802). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w25802 (cit. on p. 10).

Aksoy, B., Lusher, L., & Carrell, S. (2023). From Distraction to Dedication: Commitment Against Phone Use in the Classroom (cit. on p. 8).

Allcott, H., Gentzkow, M., & Song, L. (2022). Digital Addiction. *American Economic Review*, *112*(7), 2424–2463. https://doi.org/10.1257/aer.20210867 (cit. on p. 8).

Allcott, H., Kim, J., Taubinsky, D., & Zinman, J. (2021). Are High-Interest Loans Predatory? Theory and Evidence from Payday Lending. *The Review of Economic Studies*, *89*(3), 1041–1084. https://doi.org/10.1093/restud/rdab066 (cit. on pp. 8, 9).

Amador, M., Werning, I., & Angeletos, G.-M. (2006). Commitment vs. Flexibility. *Econometrica*, *74*(2), 365–396. https://doi.org/10.1111/j.1468-0262.2006.00666.x (cit. on p. 10).

Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over Time: Dynamic Inconsistency in Real Effort Tasks. *The Quarterly Journal of Economics*, *130*(3), 1067–1115. https://doi.org/10.1093/qje/qjv020 (cit. on pp. 6, 11, 18, 46).

Augenblick, N., & Rabin, M. (2019). An Experiment on Time Preference and Misprediction in Unpleasant Tasks. *The Review of Economic Studies*, *86*(3), 941–975. https://doi.org/10.1093/restud/rdy019 (cit. on pp. 7, 10, 18, 21, 49, 53).

Avery, M., Giuntella, O., & Jiao, P. (2019). Why Don't We Sleep Enough? A Field Experiment among College Students (cit. on p. 8).

Bai, L., Handel, B., Miguel, E., & Rao, G. (2021). Self-Control and Demand for Preventive Health: Evidence from Hypertension in India. *The Review of Economics and Statistics*, *103*(5), 835–856. https://doi.org/10.1162/rest_a_00938 (cit. on pp. 5, 7, 8, 10, 17, 35).

Bhattacharya, J., Garber, A. M., & Goldhaber-Fiebert, J. D. (2015). Nudges in Exercise Commitment Contracts: A Randomized Trial (cit. on p. 35).

Bonein, A., & Denant-Boèmont, L. (2015). Self-control, commitment and peer pressure: A laboratory experiment. *Experimental Economics*, *18*(4), 543–568. https://doi.org/10.1007/s10683-014-9419-7 (cit. on pp. 7, 9).

Brune, L., Chyn, E., & Kerwin, J. (2021). Pay Me Later: Savings Constraints and the Demand for Deferred Payments (cit. on p. 8).

Burger, N., & Lynham, J. (2010). Betting on weight loss . . . and losing: Personal gambles as commitment mechanisms. *Applied Economics Letters*, *17*(12), 1161–1166. https://doi.org/10.1080/00036840902845442 (cit. on pp. 8, 35).

Carrera, M., Royer, H., Stehr, M., Sydnor, J., & Taubinsky, D. (2022). Who Chooses Commitment? Evidence and Welfare Implications. *The Review of Economic Studies*, *89*(3), 1205–1244. https://doi.org/10.1093/restud/rdab056 (cit. on pp. 5, 7–10, 35, 47, 48).

Casaburi, L., & Macchiavello, R. (2019). Demand and Supply of Infrequent Payments as a Commitment Device: Evidence from Kenya. *American Economic Review*, *109*(2), 523–555. https://doi.org/10.1257/aer.20180281 (cit. on p. 8).

Cerrone, C., Chakraborty, A., Kim, H. J., & Lades, L. K. (2023). *Estimating Present Bias and Sophistication over Effort and Money* (Working paper No. No. 359). University of California, Department of Economics, Davis, CA. (Cit. on p. 11).

Chaloupka, F., Levy, M., & White, J. (2019, December). *Estimating Biases in Smoking Cessation: Evidence from a Field Experiment* (tech. rep. No. w26522). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w26522 (cit. on pp. 5, 7, 8, 10).

DellaVigna, S., & Malmendier, U. (2004). Contract Design and Self-Control: Theory and Evidence*. *The Quarterly Journal of Economics*, *119*(2), 353–402. https://doi.org/10.1162/0033553041382111 (cit. on p. 5).

Derksen, L., Kerwin, J., Reynoso, N. O., & Sterck, O. (2021). Appointments: A More Effective Commitment Device for Health Behaviors. Retrieved November 19, 2023, from http://arxiv.org/abs/2110.06876 (cit. on pp. 35, 37, 56).

Ek, C., & Samahita, M. (2023). Too much commitment? An online experiment with tempting YouTube content. *Journal of Economic Behavior & Organization*, *208*, 21–38. https://doi.org/10.1016/j.jebo.2023.01.019 (cit. on p. 6).

Exley, C. L., & Naecker, J. K. (2017). Observability Increases the Demand for Commitment Devices. *Management Science*, *63*(10), 3262–3267. https://doi.org/10.1287/mnsc.2016.2501 (cit. on pp. 8, 9, 33, 35).

Fedyk, A. (2016). Asymmetric Naivete: Beliefs About Self-Control. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2727499 (cit. on pp. 7, 10, 36, 49).

Galperti, S. (2015). Commitment, Flexibility, and Optimal Screening of Time Inconsistency. *Econometrica*, *83*(4), 1425–1465. https://doi.org/10.3982/ECTA11851 (cit. on p. 10).

Giné, X., Karlan, D., & Zinman, J. (2010). Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation. *American Economic Journal: Applied Economics*, *2*(4), 213–235. https://doi.org/10.1257/app.2.4.213 (cit. on pp. 5, 8, 35).

Halpern, S. D., French, B., Small, D. S., Saulsgiver, K., Harhay, M. O., Audrain-McGovern, J., Loewenstein, G., Brennan, T. A., Asch, D. A., & Volpp, K. G. (2015). Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation. *New England Journal of Medicine*, *372*(22), 2108–2117. https://doi.org/10.1056/NEJMoa1414293 (cit. on pp. 5, 8).

Heidhues, P., & Kőszegi, B. (2009). Futile Attempts at Self-Control. *Journal of the European Economic Association*, *7*(2-3), 423–434. https://doi.org/10.1162/JEEA.2009.7.2-3.423 (cit. on pp. 5, 8).

Imai, T., Rutter, T. A., & Camerer, C. F. (2021). Meta-Analysis of Present-Bias Estimation using Convex Time Budgets. *The Economic Journal*, *131*(636), 1788–1814. https://doi.org/10.1093/ej/ueaa115 (cit. on pp. 11, 46).

John, A. (2020). When Commitment Fails: Evidence from a Field Experiment. *Management Science*, *66*(2), 503–529. https://doi.org/10.1287/mnsc.2018.3236 (cit. on pp. 5, 7–9, 17, 35).

Kaur, S., Kremer, M., & Mullainathan, S. (2015). Self-Control at Work. *Journal of Political Economy*, *123*(6), 1227–1277. https://doi.org/10.1086/683822 (cit. on pp. 7–9, 31, 35, 49, 54).

Kueng, L. (2018). Excess Sensitivity of High-Income Consumers*. *The Quarterly Journal of Economics*, *133*(4), 1693–1751. https://doi.org/10.1093/qje/qjy014 (cit. on pp. 6, 9, 10).

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, *112*(2), 443–477. Retrieved August 19, 2024, from http://www.jstor.org/stable/2951242 (cit. on p. 5).

Laibson, D. (2015). Why Don't Present-Biased Agents Make Commitments? *American Economic Review*, *105*(5), 267–272. https://doi.org/10.1257/aer.p20151084 (cit. on pp. 9, 47).

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328. https://doi.org/10.1037/0022-3514.32.2.311 (cit. on p. 6).

Leary, J. B., & Wang, J. (2016). *Liquidity Constraints and Budgeting Mistakes: Evidence from Social Security Recipients* (Working paper). (Cit. on pp. 6, 9, 10).

Norcross, J. C., & Vangarelli, D. J. (1988). The resolution solution: Longitudinal examination of New Year's change attempts. *Journal of Substance Abuse*, *1*(2), 127–134. https://doi.org/10.1016/S0899-3289(88)80016-6 (cit. on p. 5).

O'Donoghue, T., & Rabin, M. (1999). Doing It Now or Later. *The American Economic Review*, *89*(1), 103–124. Retrieved January 9, 2024, from http://www.jstor.org/stable/116981 (cit. on p. 5).

O'Donoghue, T., & Rabin, M. (2001). Choice and Procrastination*. *The Quarterly Journal of Economics*, *116*(1), 121–160. https://doi.org/10.1162/003355301556365 (cit. on p. 11).

Posey, L. L., & Yavas, A. (2007). Screening equilibria in experimental markets. *The Geneva Risk and Insurance Review*, *32*(2), 147–167. https://doi.org/10.1007/s10713-007-0007-z (cit. on p. 10).

Royer, H., Stehr, M., & Sydnor, J. (2015). Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company. *American Economic Journal: Applied Economics*, *7*(3), 51–84. https://doi.org/10.1257/app.20130327 (cit. on pp. 5, 7, 8, 35).

Schilbach, F. (2019). Alcohol and Self-Control: A Field Experiment in India. *American Economic Review*, *109*(4), 1290–1322. https://doi.org/10.1257/aer.20170458 (cit. on pp. 5, 7–9, 35).

Schwartz, J., Mochon, D., Wyper, L., Maroba, J., Patel, D., & Ariely, D. (2014). Healthier by Precommitment. *Psychological Science*, *25*(2), 538–546. https://doi.org/10.1177/0956797613510950 (cit. on pp. 8, 35).

Shapira, Z., & Venezia, I. (1999). Experimental Tests of Self-Selection and Screening in Insurance Decisions. *The Geneva Papers on Risk and Insurance Theory*, *24*(2), 139–158. Retrieved August 27, 2024, from http://www.jstor.org/stable/41953375 (cit. on pp. 6, 9, 10).

Toussaert, S. [Séverine]. (2018). Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment. *Econometrica*, *86*(3), 859–889. https://doi.org/10.3982/ECTA14172 (cit. on p. 10).

Toussaert, S. [Severine]. (2019). Revealing temptation through menu choice: Field evidence (cit. on p. 8).

A APPENDIX FOR CHAPTER 1

A.1 Proofs

Lemma. *Consider a contract* $(\#, \$)$. *If*

$$-\frac{1}{\hat{\beta}}C(\#, \gamma_L) + w\# \geq -\$ - \frac{1}{\hat{\beta}}C(\widehat{t}_L, \gamma_L) + w\widehat{t}_L, \tag{1.10}$$

then

$$-\frac{1}{\hat{\beta}}C(\widehat{t}^{\star}_S, \gamma_S) + w\widehat{t}^{\star}_S \geq -\$ - \frac{1}{\hat{\beta}}C(\widehat{t}_S, \gamma_S) + w\widehat{t}_S, \tag{1.11}$$

where $\widehat{t}^{\star}_S \coloneqq \max\{\widehat{t}_S, \#\}$.

Proof. If $\# \leq \widehat{t}_S$, (1.11) trivially holds due to $\widehat{t}^{\star}_S = \widehat{t}_S$. If also holds if $\widehat{t}_S = \widehat{t}_L$ because of (1.10). Let us consider $\# > \widehat{t}_S > \widehat{t}_L$ and re-arrange (1.10):

$$\left[w\# - \frac{1}{\hat{\beta}}C(\#, \gamma_L)\right] - \left[w\widehat{t}_L - \frac{1}{\hat{\beta}}C(\widehat{t}_L, \gamma_L)\right] \geq -\$.$$

Let us denote $u(t, \gamma) \coloneqq wt - \frac{1}{\hat{\beta}}C(t|\gamma)$:

$$u(\#, \gamma_L) - u(\widehat{t}_L, \gamma_L) \geq -\$.$$

Since $\# > \widehat{t}_L$, we can re-write the LHS as a sum of marginal losses:

$$\left[u(\#, \gamma_L) - u(\# - 1, \gamma_L)\right] + \left[u(\# - 1, \gamma_L) - u(\# - 2, \gamma_L)\right] + \cdots + \left[u(\widehat{t}_L + 1, \gamma_L) - u(\widehat{t}_L, \gamma_L)\right] \geq -\$.$$

Since every term of the sum is a marginal loss, and, therefore, negative, the following holds because of $\# > \widehat{t}_S$ (we drop the terms at the end of LHS):

$$\left[u(\#, \gamma_L) - u(\# - 1, \gamma_L)\right] + \left[u(\# - 1, \gamma_L) - u(\# - 2, \gamma_L)\right] + \cdots + \left[u(\widehat{t}_S + 1, \gamma_L) - u(\widehat{t}_S, \gamma_L)\right] \geq -\$.$$

Since $\dfrac{\partial^2 C(., \gamma)}{\partial \gamma^2} > 0$, for any $t > \widehat{t}$ it holds that $\left[u(t, \gamma_L) - u(t-1, \gamma_L)\right] < \left[u(t, \gamma_S) - u(t-1, \gamma_S)\right]$. Then

$$\left[u(\#, \gamma_S) - u(\# - 1, \gamma_S)\right] + \left[u(\# - 1, \gamma_S) - u(\# - 2, \gamma_S)\right] + \cdots + \left[u(\widehat{t}_S + 1, \gamma_S) - u(\widehat{t}_S, \gamma_S)\right] >$$
$$> \left[u(\#, \gamma_L) - u(\# - 1, \gamma_L)\right] + \left[u(\# - 1, \gamma_L) - u(\# - 2, \gamma_L)\right] + \cdots + \left[u(\widehat{t}_S + 1, \gamma_L) - u(\widehat{t}_S, \gamma_L)\right] \geq -\$,$$

implying $u(\#, \gamma_S) > u(\widehat{t}_S, \gamma_S) > -\$$, and making (1.11) hold.

Proposition on Page 16

Let us assume that $\#_B = t^\star_{1S}$. Then, the expected consumption flow from the partial commitment is

$$p\left(wt^\star_{1S} - C(t^\star_{1S}|\gamma_S)\right) + (1-p)\left(w(t^\star_{1S} - \widehat{t}_S) - \frac{1}{\widehat{\beta}}\left(C(t^\star_{1S}|\gamma_S) - C(\widehat{t}_S|\gamma_S)\right)\right),$$

and from the full commitment, it is

$$p\left(w\widehat{t}_S - C(\widehat{t}_S|\gamma_S)\right) + (1-p)\left(w\#^\star_B - C(\#^\star_B|\gamma_L)\right). \tag{1.12}$$

Let us take the difference and omit the term related to the costs for Long, as we assume it to be not relevant for the change in the costs for Short. After grouping the terms by $t^\star_{1S}$ and $\widehat{t}_S$, we get

$$p\left(wt^\star_{1S} - C(t^\star_{1S}|\gamma_S)\right) + (1-p)\left(wt^\star_{1S} - \frac{1}{\widehat{\beta}}\left(C(t^\star_{1S}|\gamma_S)\right)\right) - \tag{1.13}$$

$$-(1-p)\left(w\widehat{t}_S - \frac{1}{\widehat{\beta}}C(\widehat{t}_S|\gamma_S)\right) - p\left(w\widehat{t}_S - C(\widehat{t}_S|\gamma_S)\right), \tag{1.14}$$

or

$$\left[wt^\star_{1S} - \left(p + \frac{1-p}{\widehat{\beta}}\right)C(t^\star_{1S}|\gamma_S)\right] - \left[w\widehat{t}_S - \left(p + \frac{1-p}{\widehat{\beta}}\right)C(\widehat{t}_S|\gamma_S)\right]. \tag{1.15}$$

Let us now simplify notations and define function F, so we can analyze its derivative with respect to γ.

$$F \coloneqq w\left[t^\star - \widehat{t}\right] - \left(p + \frac{1-p}{\widehat{\beta}}\right)\left(C(t^\star|\gamma) - C(\widehat{t}|\gamma)\right), \tag{1.16}$$

where $t^\star = (C')^{-1}(w|\gamma)$ and $\widehat{t} = (C')^{-1}(\widehat{\beta}w|\gamma)$.

$$\frac{\partial F}{\partial \gamma} = w\left(\frac{\partial t^\star}{\partial \gamma} - \frac{\partial \widehat{t}}{\partial \gamma}\right) - \left(p + \frac{1-p}{\widehat{\beta}}\right)\left[\left(\frac{\partial C}{\partial \gamma}(t^\star) + \frac{\partial C}{\partial t^\star}\frac{\partial t^\star}{\partial \gamma}\right) - \left(\frac{\partial C}{\partial \gamma}(\widehat{t}) - \frac{\partial C}{\partial \widehat{t}}\frac{\partial \widehat{t}}{\partial \gamma}\right)\right],$$

$$\frac{\partial F}{\partial \gamma} = w\left(\frac{\partial t^\star}{\partial \gamma} - \frac{\partial \widehat{t}}{\partial \gamma}\right) - \left(p + \frac{1-p}{\widehat{\beta}}\right)\left[\left(\frac{\partial C}{\partial \gamma}(t^\star) - \frac{\partial C}{\partial \gamma}(\widehat{t})\right) + w\left(\frac{\partial t^\star}{\partial \gamma} - \widehat{\beta}\frac{\partial \widehat{t}}{\partial \gamma}\right)\right],$$

$$\frac{\partial F}{\partial \gamma} = w \underbrace{\left[\left(\frac{\partial t^\star}{\partial \gamma} - \frac{\partial \widehat{t}}{\partial \gamma} \right) - \left(\frac{\partial t^\star}{\partial \gamma} - \widehat{\beta} \frac{\partial \widehat{t}}{\partial \gamma} \right) \left(p + \frac{1-p}{\widehat{\beta}} \right) \right]}_{A} - \left(p + \frac{1-p}{\widehat{\beta}} \right) \underbrace{\left[\frac{\partial C}{\partial \gamma}(t^\star) - \frac{\partial C}{\partial \gamma}(\widehat{t}) \right]}_{B}. \tag{1.17}$$

$$\frac{\partial t^\star}{\partial \gamma} = -\frac{\frac{\partial^2 C}{\partial t \partial \gamma}(t^\star)}{\frac{\partial^2 C}{\partial t^2}(t^\star)}, \qquad \frac{\partial \widehat{t}}{\partial \gamma} = -\frac{\frac{\partial^2 C}{\partial t \partial \gamma}(\widehat{t})}{\frac{\partial^2 C}{\partial t^2}(\widehat{t})}. \tag{1.18}$$

- We argue that $A > 0$. First, $\frac{\partial t^\star}{\partial \gamma} < 0$ and $\frac{\partial \widehat{t}}{\partial \gamma} < 0$. Their difference depends on the proportion between $\frac{\partial^3 C}{\partial t^2 \partial \gamma}$ and $\frac{\partial^3 C}{\partial t^3}$. We argue that, in our setup, $\frac{\partial^3 C}{\partial t^2 \partial \gamma}$ is larger than $\frac{\partial^3 C}{\partial t^3}$ (they are both positive). The expression $\frac{\partial^2 C}{\partial t \partial \gamma}$ can be interpreted as how fast participants get tired as they get more letters in each task. The expression $\frac{\partial^2 C}{\partial t^2}$ can be interpreted as how fast participants get tired as they go from task to task. $\frac{\partial^2 C}{\partial t \partial \gamma}$ noticeably increases in t: by the time the participants solve many tasks, even a marginal increase in t accumulates. Meanwhile, we can expect $\frac{\partial^2 C}{\partial t \partial \gamma}$ to barely change in t over time.[18] Since $\frac{\partial^3 C}{\partial t^2 \partial \gamma} > \frac{\partial^3 C}{\partial t^3} > 0$, $\frac{\partial t^\star}{\partial \gamma} - \frac{\partial \widehat{t}}{\partial \gamma} < 0$, $(\frac{\partial t^\star}{\partial \gamma} - \widehat{\beta} \frac{\partial \widehat{t}}{\partial \gamma}) \left(p + \frac{1-p}{\widehat{\beta}} \right) < \frac{\partial t^\star}{\partial \gamma} - \frac{\partial \widehat{t}}{\partial \gamma} < 0$, and $A > 0$.

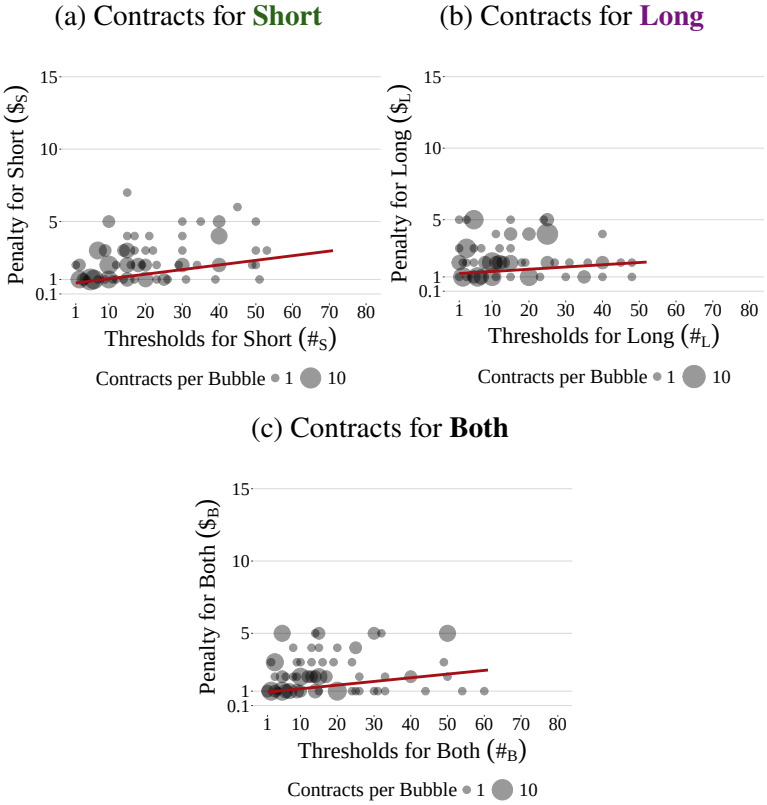
- Further, we argue that B is close to 0. First, $B > 0$ as $\frac{\partial^2 C}{\partial t \partial \gamma} > 0$. However, we believe that $\frac{\partial^2 C}{\partial t \partial \gamma}$ is very close to 0. That is because $\frac{\partial C}{\partial t}$ is the effort required for doing one additional task, and increasing the number of letters barely changes it: the keyboard remains the same compared to a new task, for example.

- Since $A > 0$ and we can consider B sufficiently close to 0, $\partial F / \partial \gamma > 0$.

[18] We plan to check our arguments empirically.

A.2 Main Results on Day 2

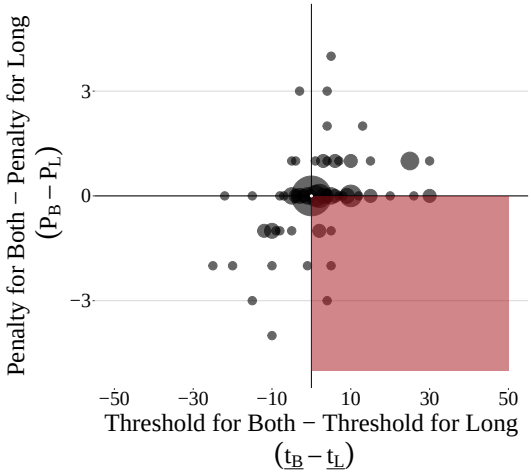
Figure A.1: Commitment Take-up by Contract Type (Day 2, recreating Figure 1.5)



Note: Each panel displays the relationship between penalty and threshold levels for a different type of commitment contract: (a) contracts that apply only to Short tasks, (b) contracts that apply only to Long tasks, and (c) contracts that apply to both task types. In each plot, one bubble represents a group of similar contracts; the bubble size reflects the number of such contracts (with a maximum of 10 per bubble). The red line indicates the linear correlation between penalty size and threshold level.

Figure A.2: Partial Contracts, %: **29.7 [18.1, 41.3]** (Day 2, recreating Figure 1.6)

$$\begin{bmatrix} \#_B - \#_L > 0 \wedge \$_B - \$_L \leq 0, & \text{or} \\ \#_B = \#_L \wedge \$_B - \$_L < 0 \end{bmatrix}$$



Note: The figure displays the commitment take-up, where red-shaded areas delineate the regions associated with partial contracts. The subtitles give the proportion of contracts for Both, which are identified as partial under the corresponding definition, along with a 95% CI with standard errors clustered by individuals. One observation is a contract for Both; a bubble represents a group of similar contracts for Both. A larger bubble corresponds to a larger number of similar contracts. The x-axis denotes the difference between the threshold in the contract for Both and the threshold in the contract for Long under the same piece-rate wage ($\#_B - \#_L$). The y-axis denotes the difference between the penalty in the contract for Both and the penalty in the contract for Long under the same piece-rate wage ($\$_B - \$_L$). The details of the partial contract definitions are provided below the subtitles. The partial contracts are identified among the 200 contracts for Both that have thresholds above 1. We omit 11 out of 200 contracts from the figure as outliers.

A.3 Planned Failure Capacity

Note: All notations here are introduced exclusively for A.3 and are unrelated to the other parts of the text. We do so to keep the notations simple.

We build a simple model of penalty-based commitment that puts little to no restriction on the distribution of costs.[19] We build separately from the model in the main text as it very general. Its goal is to show a tractable explanation for why

[19]The setup and notations we use in the analysis below are not related to that in the main text.

we consider the chance commitment makes a difference to be the maximum rate of planned failures.

We consider a two-period model for an agent who can take up a penalty-based commitment device at zero cost. Any device of this kind has a penalty of size p, which the agent loses in period 2 if she does not do the action. The device is available in period 1, and it can affect whether she performs an action ($a = 1$) in period 2 or abstains ($a = 0$). In period 1, the agent prefers her future self to perform the action as long as her benefits exceed the discounted costs. In our model, we will assume that the benefits are not discounted regardless of the period. This assumption does not affect the results and aligns with studies emphasizing that the present bias for action is much more acute than for money (Augenblick, Niederle, and Sprenger, 2015; Imai et al., 2021, Cherrone-Chakraborty-Kim-Lades WP). Thus, in period 1, the agent wants her future decision rule to be

$$a = 1 \quad \Leftrightarrow \quad b - c \geq 0, \tag{1.19}$$

where b represents the benefits and c represents the discounted costs.

Suppose that the agent believes she has a present bias of size $\widehat{\beta} \in (0, 1)$. In period $t = 1$, she thinks that her action in period 2 will be determined by the following decision rule:

$$\widehat{a} = 1 \quad \Leftrightarrow \quad b - \frac{1}{\widehat{\beta}} c \geq 0, \tag{1.20}$$

where $\dfrac{1}{\widehat{\beta}} c$ represents the immediate costs.

We will assume that delayed costs c are stochastic and distributed as follows:

1. With chance $\phi \in (0, 1)$, $c \leq \widehat{\beta} b$: the agent does the action no matter whether she takes up commitment or not.

2. With chance $\xi \in (0, 1 - \phi]$, $c \in (\widehat{\beta} b; b]$: despite the agent in period 1 wanting the action to be performed, she will not do it unless she takes up commitment. Let us assume that, conditional on $c \in (\widehat{\beta} b; b]$, c has an expected value of μb, where $\mu \in \left(\widehat{\beta}, 1\right)$.

3. With chance $1 - \phi - \xi$, $c \to +\infty$: in period 1, the agent wants her future self to abstain from the action, and she will not do it no matter whether she takes up commitment or not. Making this realization of c extremely large

is not necessary; we only want to emphasize that there is no need to restrict the value of a realization of costs as long as the agent knows she would pay a penalty upon this realization. The possibility of c being extremely large with a non-moderate chance is missing from the analysis by Laibson, 2015 and Carrera et al., 2022, as they consider the distribution as a whole to be determined by a differentiable cumulative distribution function. From the empirical perspective, the possibility that $c \to +\infty$ can be incorporated as a positive chance with which the agent pays the penalty no matter the realization of costs, which can be assumed to otherwise have a regular distribution (for example, exponential).

The costs are realized in period 2 right before the agent decides if she does the action. Therefore, in period 1, the agent decides whether to take up commitment or not based on its distribution rather than a deterministic value.

We will now consider a penalty-based commitment contract with penalty $p = \left(\frac{1}{\widehat{\beta}} - 1\right) b$: this contract will make the agent choose $a = 1$ if and only if it is what she prefers her future self to do in period 1. We call such device *benefit-based*. We use it as an example to illustrate that the demand for commitment with *some* penalty is not fully eroded by a large chance of failure.

Proposition. *The agent takes up a benefit-based commitment device if and only if:*

$$\xi(1-\mu) - (1-\xi-\phi)\left(\frac{1}{\widehat{\beta}} - 1\right) \geq 0. \tag{1.21}$$

Proof. If the agent takes up this contract in period 1, she believes her expected payoff to be as follows:

$$\xi(b-\mu b) + (1-\phi-\xi)(-p) = \left[\xi(1-\mu) - (1-\xi-\phi)\left(\frac{1}{\widehat{\beta}} - 1\right)\right] b. \tag{1.22}$$

The statement of the proposition follows from the assumption that the agent makes her take-up decision based on the expected payoff.

Expression (1.21) can be positive under reasonable parameters. For example, the empirical estimates from the study by Carrera et al., 2022 allow for $\phi = 0.22$, $\xi = 0.43$, $\widehat{\beta} = 0.84$. Then, (1.21) is positive if μ is negligibly larger than $\widehat{\beta}$, which stands for the contract to be useful under the lowest possible costs. Although it is

unlikely to be the case, it would allow all failures (35%) to be planned. Planned failure can still be as high as 31% of commitment take-up with a more realistic $\mu = 0.87$.[20]

Corollary 1 *The agent does not take up the benefit-based contract if the chance that benefits exceed the delayed costs is larger than the chance the agent in period* 1 *prefers her future self to do the action, while she would not do it without commitment.*

Corollary 1 follows from rearranging the terms in expression (1.21) and repeats what was established by Carrera et al., 2022 under arguably less restrictive assumptions on the cost function.

Corollary 2 *A commitment device can be in demand even if the condition in Corollary 1 does not hold.*

Corollary 2 follows from considering the distribution of c such that the chance c is close to b is extremely small. If it is zero, then the expected costs while $c \in (\widehat{\beta}b; b]$ remain μb, while the penalty smaller than $\left(\frac{1}{\widehat{\beta}} - 1\right) b$ can be selected, thereby increasing expression (1.21). Carrera et al., 2022 might not consider this possibility as their assumptions make expected benefits from commitment increase in p so long as c remains strictly below b. It cannot be considered an omission as such a distribution of costs does not seem likely to be empirically observed, and therefore Corollary 2 is merely an exercise.

A.4 General Data Description

The 69 participants in our sample were offered 1, 242 contracts one week ahead and the same amount thirty minutes ahead. Participants accepted 59.4% of the contracts one week ahead and 51.3% thirty minutes ahead. These rates are among the highest for penalty-based studies,[21] likely due to the flexibility of each contract. These rates

[20] While emphasizing that the chance of costs exceeding benefits needs to be moderate, Carrera et al., 2022 focus on the distribution of costs of going to the gym. We abandon this approach as their commitment devices are imposed on, for example, going to the gym 12 or more times. Notably, their analysis fully applies to the piece-rate incentives that they examine both theoretically and empirically in great detail.

[21] Usually between 10% and 70%, (see Carrera et al., 2022, for a summary).

also align with the proportion of people aware of their present bias, as observed in the studies by Augenblick and Rabin, 2019 and Fedyk, 2016.[22]

In our setup, commitment take-up (as a binary decision) is an individual characteristic and depends almost exclusively on whether the individual is aware of their present bias. That is because we allow for a very flexible contract, which is strictly beneficial as long as the individual expects herself to do fewer tasks in the future than she would prefer presently. Indeed, the take-up one week ahead (Figure A.3a) shows little different in take-up across wages and contract types. However, thirty minutes ahead, there is a tendency for higher take-up for larger wages. There also seems to be a tendency for contracts for Short to be in more demand than contracts for Long. One possible explanation is that, if costs are too high or the wage is too low, a small extent of present bias ($\widehat{\beta} \approx 1$) can lead to no difference between the number of tasks that the agent wants herself to do in the future and expects herself to do in the future, as the number of tasks is discrete.

Further, we find most participants can be classified into those who commit and those who do not. One week ahead, 51 out of 69 participants took up at least 15 or at most 3 contracts out of 18. Thirty minutes ahead, this number was 50.

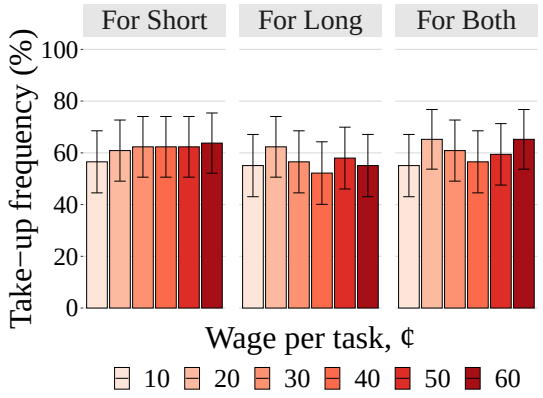
In addition to the take-up rates on contract level, we consider the take-up rates on individual level. Out of 69 participants, 35 (46.3%) took up at least 15 contracts out of 18 one week ahead. Thirty minutes ahead, this number is 25 (36.2%), and 19 participants (27.5%) took up at least 15 contracts each day. We attribute the distinction between the days to the small time difference between thirty minutes ahead and the time on performing the tasks: the participants might see their interests mostly aligned with those of their future selves, and, therefore, see no need in a contract. Another possibility is that they had costly take-up experimentation one week ahead, and did not benefit from it thirty minutes ahead (Kaur et al., 2015). We discuss whether take-up stems from awareness of present bias or something else in the corresponding section.

Result 1. The decision to commit is largely unaffected by the difficulty of the tasks or the wage received by the participant.

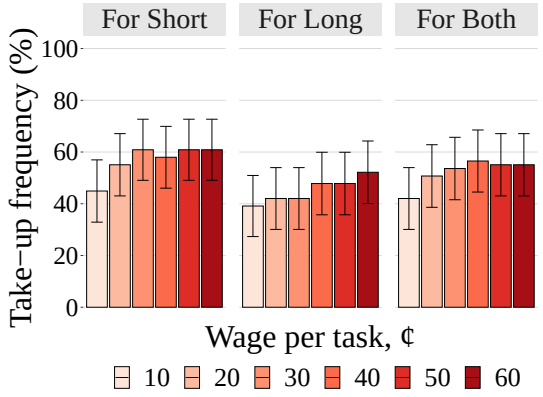
[22] About 40–45% in the study by Fedyk, 2016 (based on the kernel density graph, the median is 1) and 54–60% in the study by Augenblick and Rabin, 2019.

Figure A.3: Commitment take-up by wage

(a) One week ahead



(b) Thirty minutes ahead

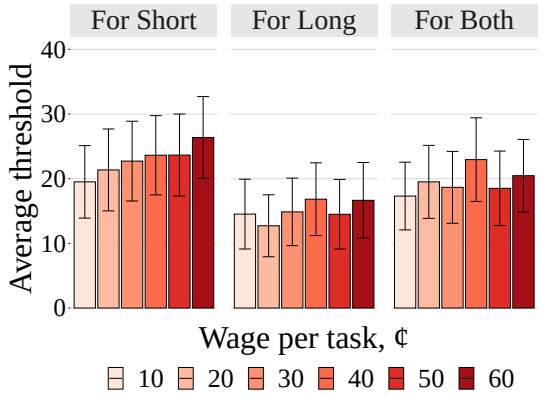


Thresholds

Figure A.4 shows the average thresholds by wage across all taken-up contracts. First, the pairwise differences between the average thresholds for Short and for Long under the same wage are only occasionally statistically significant one week ahead and never thirty minutes ahead. Second, there is scant evidence of any impact of wages on average thresholds: higher piece-rate wages have a negligible impact on the thresholds on either day. Even in contracts for Short, the difference in average thresholds between the lowest piece-rate wage of 0.10 and the highest wage of 0.60 is not significant. Similarly, penalties, on average, do not increase with wage. Further for each participant who signed up for at least two different wages in the contract for Short one week ahead, we collected the thresholds for the lowest and highest wages they signed up for and found no statistically significant difference between these thresholds. Likewise, no difference emerges in the average thresholds between those who signed up for the minimum and maximum wages.

Figure A.4: Average threshold by wage

(a) One week ahead



(b) Thirty minutes ahead

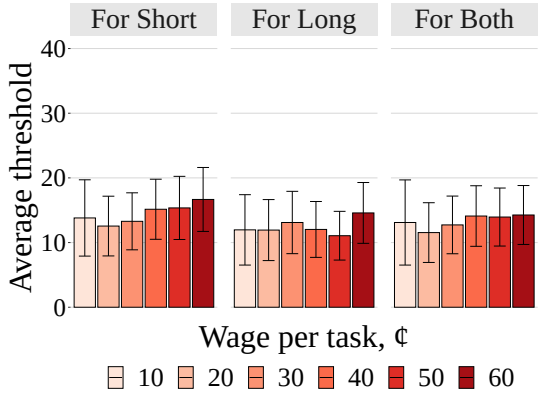
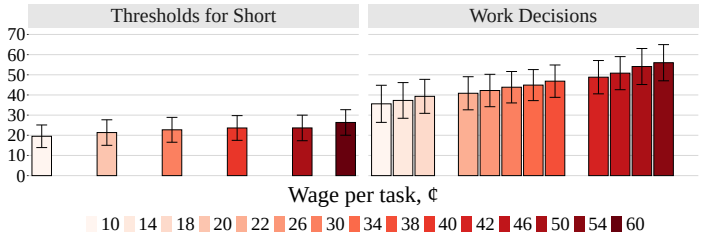
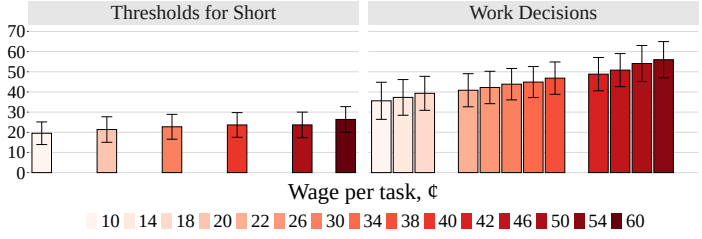


Figure A.5: Thresholds vs. Work Decisions

(a) One week ahead



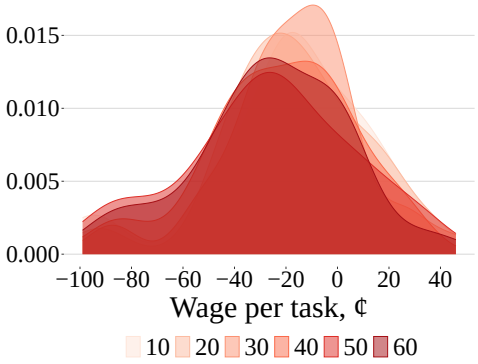
(b) Thirty minutes ahead



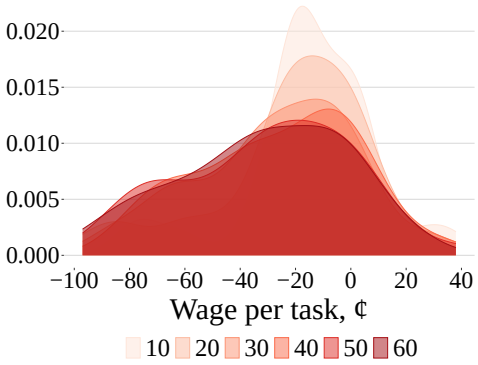
Now let us demonstrate that the thresholds for Short tasks are, on average, significantly lower than the corresponding work decisions. We begin this comparison by analyzing how these thresholds align with the participants' responses about how many tasks they intended to complete. The analysis starts with the sample of 38 participants one week ahead and 36 thirty minutes ahead, each of whom took up 5 out of 6 offered contracts.[23] Figure A.5 shows that the thresholds in their contracts for Short tasks are, on average, significantly lower than their declared work decisions at corresponding wages.

Figure A.6: Thresholds minus Work Decisions

(a) One week ahead



(b) Thirty minutes ahead



We further compare thresholds and work decisions on an individual level. Figure A.6 illustrates the distributions of the differences between thresholds for Short tasks and corresponding Work Decisions.[24] Further, more than half of the thresholds are significantly lower than work decisions.

[23] Most participants who accepted exactly 5 contracts out of 6 typically did not take up the contract offering the minimum wage of 10¢.

[24] Work decisions under 10, 18, 30, 48, 50, and 54 cents piece-rate wages corresponding to contracts under piece-rate wages of 10, 20, 30, 40, 50, and 60 cents, respectively.

Figure A.6 suggests that more than half of the participants choose commitments that do not encourage them to complete more tasks than they expect to do without commitment. Given our discovery that commitment appears to be an individual characteristic, we explored whether the difference between threshold and work decision might also be a personal trait. To this end, we revisited our sample of participants who took up at least five contracts for Short out of six. We calculated the average difference between the threshold and work decision for each participant. Although each comparison involves only five or six observations, one week ahead (thirty minutes ahead) 16 (18) out of 38 (36) participants exhibited thresholds lower than work decisions at a 0.05 significance level, and 20 (24) at a 0.1 significance level. Meanwhile, only one participant had thresholds significantly higher than work decisions on average one week ahead, with a p-value of 0.06 at the 0.1 level. There were no such participants thirty minutes ahead.

We suggest two interpretations for the difference between work decisions and commitment thresholds. Our first interpretation is that our split between Short tasks and Long tasks does not account for all the uncertainty of costs. Specifically, under uncertain costs, a participant needs to choose a work decision that is suboptimal for many cost realizations. Imagine a distribution of costs that is skewed towards lower values but has a large span. The participant might set her work decision high as she wants to benefit from the lower realizations of costs. Meanwhile, a commitment keeps the participant capable of doing more tasks than the threshold. Therefore, she might set the threshold low to avoid the consequences of extremely low costs but still reap the benefits of low-cost realizations.

Let us briefly discuss how this interpretation affects our identification strategy. Suppose that a participant takes up commitment for Long and the costs of doing Long tasks are uncertain. Then there are cost realizations for which penalty P_L in commitment for Long is the minimum penalty that ensures reaching threshold $\overline{e}_L$. Then a partial contract will make the individual fail for these cost realizations. Therefore, our identification approach is still indicative of an intention to fail with a positive chance.

Our second interpretation suggests that many participants miss an opportunity to boost their performance with commitment altogether. If they prefer commitment per se, which is consistent with Result 1, they can choose a 'safe' threshold they believe they would reach even without commitment. This result aligns with findings from Augenblick and Rabin, 2019, who observed that despite having the opportunity,

participants do not leverage tasks as a commitment device. Further, Kaur et al., 2015 suggest that participants might engage in *costly experimentation*, which, again, can make one take up a commitment without an intent to use it. If this interpretation prevails, our identification of planned failure is flawed.

These two interpretations are not mutually exclusive. Unfortunately, our data does not allow us to distinguish clearly between them. Notably, the uncertainty-based interpretation is consistent with the approach to commitment as a way for people to enhance their performance. Meanwhile, the non-instrumental interpretation requires developing a new perspective on how (and if) commitment options should be offered.

Result 2. On average, thresholds for Short tasks are lower than the thresholds for Long tasks.

Result 3. At least half of the participants choose a commitment threshold for Short significantly lower than our prediction.

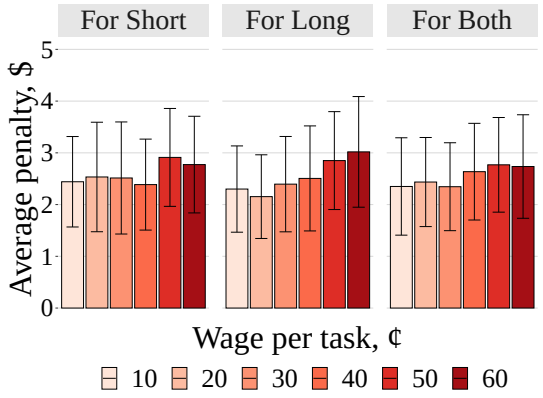
Before we proceed, let us briefly explain how the two interpretations affect our identification of partial commitments. We assume that the participants choose the minimum penalty that allows them to follow through with commitment whenever they plan. Then setting a threshold to be higher or a penalty to be lower

Penalties

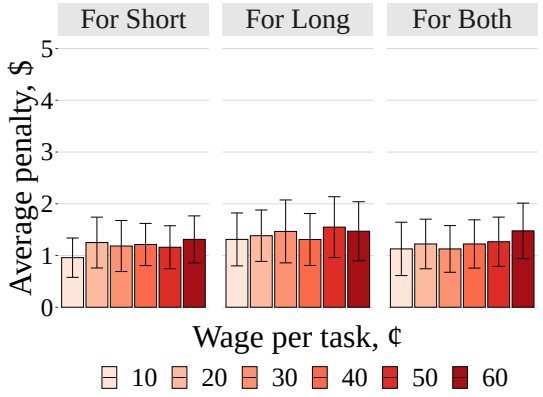
We summarize the average penalties in Figure A.7. Similarly to the size of thresholds, the size of penalties does not change across the types of contracts and wages. Meanwhile, the average penalties are significantly smaller thirty minutes ahead than one week ahead. With the thresholds for Long being approximately the same on Days 1 and 2 (see Figure A.4), it corresponds to $1 > \widehat{\beta}^{\text{Day2}} > \widehat{\beta}$. Further, the small size of the penalty thirty minutes ahead is consistent with $\widehat{\beta}^{\text{Day2}}$ being close to 1.

Figure A.7: Average penalty by wage

(a) One week ahead



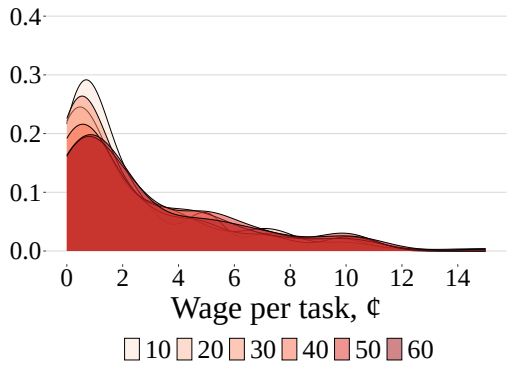
(b) Thirty minutes ahead



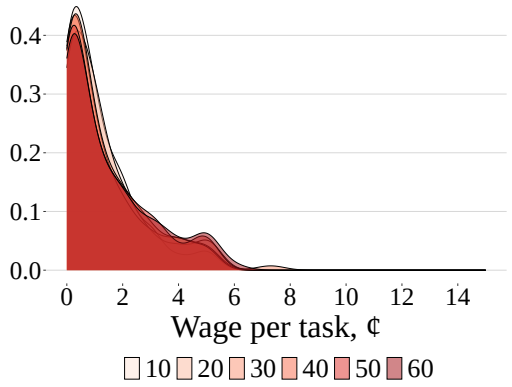
Despite that the data on penalties is mostly consistent with the predictions of the model, it is also consistent with the idea that the participants do not use commitment to boost their performance. That is because most penalties are small: across all commitment contracts, 49% and 62% of penalties do not exceed $1 one week ahead and thirty minutes ahead, respectively. Figure A.8 shows the distribution of penalties by wage for both days; we pool all three types of contracts together as there is little observable difference in the distributions of their penalties.

Figure A.8: Distribution of penalties by wage

(a) One week ahead



(b) Thirty minutes ahead



We have two interpretation for the small size of most penalties. The first interpretation is consistent with the studies that find people mostly unaware of their present bias. With $\widehat{\beta}$ only slightly below 1, they recognize that they require a commitment contract but believe that a small penalty would be enough for achieving their goal. Our second interpretation suggests that choosing *some* penalty is enough for the participant as they value following through per se. This interpretation is consistent with the study by Derksen et al., 2021, who find that doctor appointments are enough for boosting the visits, and the participant need no penalty for achieving their goal. No matter the interpretation, we can conclude that, when given a choice, the participant prefer a low penalty, and therefore signing up for a high penalty is likely being due to the preference for commitment take-up without regard to commitment parameters.

Let us summarize our results for commitment decisions before we proceed to analyze planned failures. Firstly, we find that the parameters of a contract have little impact on take-up decisions: most individuals either take up (almost) all available contracts

or barely take up any. Secondly, we find that our model does not predict the choice of thresholds accurately: individuals rarely choose thresholds at least as high as the model predicts. In contrast, many participants chose a lower threshold. Finally, most individuals choose penalties that are negligible in size. This evidence suggests that most committed participants do not aim to boost their performance. Combining it with extensive demand for commitment suggests that commitment take-up requires other modeling approaches.

Result 3. Most penalties are negligible.

Chapter 2

THE IDENTIFIED HELPER EFFECT ON THE FREQUENCY OF ASKS

1 INTRODUCTION

People in need often have opportunities to ask others for help, but frequently choose not to. Children avoid asking for help in classrooms (Ryan et al., 1998), professionals refrain from seeking help from their colleagues and employees (F. Lee, 1997), and many patients in need of a kidney transplant never reach out to potential living donors (Jaroszewicz et al., 2022). Still, asking can be a crucial step in receiving help. First, potential helpers must be made aware of the need. Second, existing research highlights the power of the ask itself: communication from the recipient to the helper significantly increases giving (Andreoni and Rao, 2011).

While much of the literature explores psychological and social barriers to asking, there is a growing need for practical tools that help overcome this reluctance. Our study contributes to this effort by offering a simple and scalable approach to encouraging asks—one that can be applied across a range of contexts to help ensure more needs are recognized and met.

We examine whether changes in the identifiability of potential helpers can influence asking behavior. Identifiability has previously been shown to affect decision-making in charitable giving. For example, the well-established identifiable helper effect—where people are more likely to help individuals they know something about, rather than anonymous or statistical others—has been shown to promote prosocial behavior (we survey this and other related literature in Section 1.1). In addition, Chen and Gao (2022) find that merely knowing a donor's identity (such as through an uninformative name) encourages recipients to act more consistently with the broader goals of the donation.

Building on these findings, we investigate whether a *very weak* form of helper identifiability can influence asking behavior. Specifically, we assign uninformative ID numbers to potential helpers and reveal these numbers to those who can ask. A related subtle manipulation was shown by Small and Loewenstein (2003) to increase giving in dictator games.

In our experiment, we randomly assign roles of askers and helpers to the participants. We introduce the need through the difference in the money endowments: helpers receive more money at the beginning of the experiment. An asker begins with a negligible sum and can request help, which would be an increase in his bonus payment at the expense of one of the helpers. Some askers receive the ID number of a helper to whom they can address their request (Treatment ID). Other askers do not receive an ID; instead, their request is directed to randomly selected helpers (Treatment No ID).[1] Since our design aims to isolate the identifiability effect, the treatments are otherwise identical. Notably, the helpers in the two treatments receive the same instructions.

We find that askers who are shown the ID of their potential helper are significantly more likely to ask for help (76.5% vs. 67%). While studying the mechanism behind the change, we find a striking difference in the role of beliefs across treatments. In Treatment ID (with identified helpers), the perceived expected payoff from asking helps predict the decision to ask. In contrast, in Treatment No ID (with unidentified helpers), the expected payoff is entirely uncorrelated with asking behavior. This suggests that other factors must be contributing to the decision to ask in Treatment No ID, to the extent that they overshadow the role of expected payoff. We identify one such factor—the perceived likelihood that other askers will request help—and show that it has a strong, positive causal effect on asking in Treatment No ID.

Our study contributes to the literature in several ways. First, the simple tool for changing the frequency of asks that we find contributes to the growing literature on the receivers of help (Bénabou et al., 2025; Chen and Gao, 2022; Jaroszewicz et al., 2022; Nadler, 2015). Second, our analysis of beliefs suggests an alternative approach to studying the mechanism of the identifiable victim effect (and other identification-based effects[2]), which has so far been approached using primarily psychological (Jenni and Loewenstein, 1997; S. Lee and Feeley, 2018) and neurological (Genevsky et al., 2013; Zhao et al., 2024) methods. Third, we add to the studies that demonstrate the effect of as minor identifiability change as uninformative numbers (Aimone and Houser, 2012; Small and Loewenstein, 2003). In addition, while the previous studies were run in laboratory settings, we demonstrate the effect on the Prolific online platform.

Promoting the efficiency of charitable, fundraising, and similar organizations re-

[1] We run the two treatments separately: the askers who receive an ID do not know that there are askers who do not receive it, and vice versa.

[2] For example, Song et al., 2022 study customer identification in the service context.

quires careful attention to the frequency of asks. We show that a marginal change in the identifiability of potential helpers can lead to a substantial increase in asking behavior. The minimal nature of this intervention suggests that the result may have broad applicability, especially once the underlying mechanisms are better understood.

The paper proceeds as follows. Section 2 presents the baseline study, which demonstrates the positive effect of identifiability on the frequency of asks. Section 4 describes the follow-up sessions, which explore the role of beliefs and confirm that askers perceive no difference between helpers in the two treatments. Section 5 discusses the results and concludes with a discussion of their broader applications.

1.1 Connection to the literature

Empirical studies on recipients of help explore a variety of contexts, including, broadly defined, charitable giving (Chen and Gao, 2022; Jaroszewicz et al., 2022), advice-seeking (Chandrasekhar et al., 2018, October; F. Lee, 1997; Ryan et al., 1998), and the take-up of social benefits (Bhargava and Manoli, 2015). Some of these studies provide empirical evidence on the extent to which people refrain from seeking help when they need it. In general, this extent is difficult to estimate: not every participant in an advice-seeking study requires advice, while studies solely on those who seek help leave those who do not seek help unobserved. Jaroszewicz et al. (2022) find that approximately one-quarter of patients in need of a kidney transplant never reach out to potential living donors. Further, using data from the 2005 tax year, Bhargava and Manoli (2015) estimate that 25% of those eligible for the Earned Income Tax Credit do not claim it, with a typical non-claimant forgoing an amount equivalent to more than a month of income. These findings underscore the importance of increasing the frequency of asks, particularly in nonprofit and health-related domains.

Many studies on help-seeking behavior focus on why individuals might refrain from asking for help—such as revealing incompetence (F. Lee, 1997; Ryan et al., 1998), incorrect beliefs about help[3] (Bohns, 2016), and, broadly, psychological pain (Bénabou et al., 2025; Bohns and Flynn, 2010; Jaroszewicz et al., 2022). The variety of possible explanations suggests that overcoming moderate rates of asking may require context-specific approaches. In contrast, our study offers a simple and

[3]We thank Ania Jaroszewicz for generously sharing an early draft of her work on the impact of helpers' beliefs about the askers' desire for help and second-order beliefs of askers.

scalable tool for encouraging more people to ask for help—one that can be applied across a wide range of settings, ultimately enabling more needs to be met.

The primary motivation for exploring the effect of identifiability on help-seeking behavior is rooted in a well-established finding: people are more inclined to help when the *person in need* is identifiable. This phenomenon, widely known as the *identifiable victim effect* (see S. Lee and Feeley, 2016, for a review), captures the tendency for individuals to respond more generously when the recipient is presented as a specific, identifiable individual rather than as a vague or statistical figure. Fundraising platforms such as GoFundMe strategically leverage this effect by encouraging users to share highly personal stories, thereby increasing donations. Proposed explanations for the identifiable victim effect include the perception that a higher proportion of people in need receive help if victims are identifiable (Jenni and Loewenstein, 1997), as well as differential emotional responses to identifiable versus statistical victims (S. Lee and Feeley, 2018; Small, 2015, August). Our study contributes to this literature by showing that identification also alters decision-making conditional on beliefs. Specifically, we demonstrate that in the absence of helper identification, the motivation to maximize expected payoff is entirely overshadowed by other considerations.

There are several factors that may cause a helper's identifiability to reduce the likelihood of asking, thereby necessitating the generalization of helpers to promote it. Individuals may feel more apprehensive about being rejected by a specific, identified person than by an anonymous or statistical helper (Bénabou et al., 2025). Additionally, people tend to be more optimistic about uncertain outcomes than those already determined, making them more willing to seek help when the helper remains anonymous (Brun and Teigen, 1990; Rothbart and Snyder, 1970; Strickland et al., 1966). At the same time, social preferences can shape the effect of identifiability in competing ways. While some may avoid imposing a burden on an identifiable individual, others may perceive asking as a way to address fairness concerns (Andreoni, Aydin, et al., 2020). Whether identifiability increases or decreases asking behavior thus depends on which of these considerations dominates.

While the change in helper identifiability we introduce is minimal, it aligns with prior findings that even slight modifications to identifiability can influence decision-making. Small and Loewenstein (2003) find that simply providing a dictator with an ID for their counterpart in a dictator game increases sharing. Similarly, Aimone and Houser (2012) demonstrate that the possibility of personal betrayal discourages trust

in an investment game, isolating its personal nature by contrasting a participant's counterpart with the trustees of other participants. Our study provides further evidence that even minimal changes in identifiability can shape individual behavior. Focusing on such minimal changes is particularly important, as they are easy to implement and avoid the framing effects introduced by more detailed information.

2 BASELINE EXPERIMENT

2.1 Experiment Design

Participants in our experiment are randomly assigned to one of two roles: askers (Group A) or helpers[4] (Group B), and this randomization is common knowledge. Each helper is assigned a unique but uninformative ID number. Participants are then split equally into two treatment groups, within which askers and helpers are matched.[5]

- **Treatment ID (Identified Helpers)**: Askers are matched with their helpers before the experiment begins and learn their helper's ID number at the start of the experiment.
- **Treatment No ID (Non-Identified Helpers)**: Askers are matched with their helpers after the experiment concludes, meaning askers do not know their helper's ID number during the experiment.

Figure 2.1[6] gives the details on the difference in the two treatments in the instructions

[4] In the main text, we focus on the behavior of askers. We summarize the instructions and responses of helpers in Appendix B.3.

[5] Since our primary interest lies in askers' behavior, we assign five times more askers than helpers. Only 20% of askers are randomly selected to be matched with a helper. As a result, each matched asker interacts with exactly one helper, ensuring that, for all practical purposes, askers and helpers can be considered paired.

[6] Note to Figure 2.1: This figure illustrates the differences in instructions given to askers across treatment conditions. It also elaborates on our assignment of five times more participants to be askers than helpers in both treatments, which we do to expand the sample of askers while efficiently managing experimental costs. Since our primary interest lies in askers' behavior, only 20% of them are randomly selected to be matched with a helper. Each subfigure presents the portion of the instructions that askers receive between the role assignment and the explanation of actions and payments. In **Treatment ID**, askers are explicitly assigned a counterpart and receive their helper's unique ID number, as shown in Figure (a). Later in the instructions, they are reminded that *only* the designated Group B participant (identified by the assigned ID) can affect their payment for the experiment. Word NEXT depicts a button that the askers had to press for the new pieces of instructions to appear on the screen. In **Treatment No ID**, participants are not assigned a specific counterpart. Further in the instructions, they are reminded that *any* participant from Group B could influence their payment.

for askers. The instructions for helpers remain the same across treatments, ensuring that any observed differences in behavior stem solely from the askers' side. To ensure that askers in different treatments perceive the helpers similarly, we provide askers with a portion of the helpers' instructions. In Section 4, we discuss the two additional experiments we ran to ensure that askers perceive no difference between helpers across the two treatments.

For both askers and helpers, the experiment consists of two rounds, one of which is randomly selected to determine participants' additional payments for the entire experiment. In Round 1, askers and helpers participate in the asking game. In Round 2, they are asked to submit their beliefs about help and about other participants.

Round 1: Asking Game

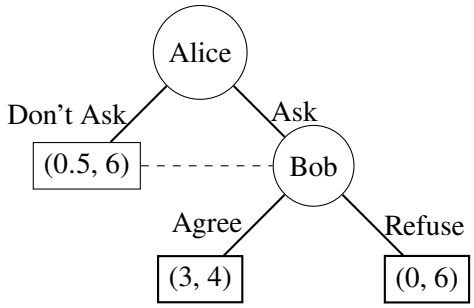
Round 1 is the main task of this experiment. For brevity, we will use the generic names Alice and Bob for the asker and the helper, respectively.

At the beginning of Round 1, Alice has a *potential bonus* of \$0.5. The potential bonus represents the additional earnings a participant can receive beyond the fixed participation payment, depending on their own and their counterpart's choices. Bob, in turn, has a potential bonus of \$6.

Alice can choose to ask Bob for help. If she does not ask, the potential bonuses remain unchanged: Alice keeps \$0.5, and Bob keeps \$6. If Alice asks, Bob then decides the outcome for both potential bonuses.

- If Bob agrees to help, Alice's potential bonus increases to \$3, while Bob's decreases to \$4.
- If Bob refuses to help, Alice's potential bonus drops to \$0, and Bob retains his \$6.

The diagram below shows the actions and gives the potential bonuses in parentheses (Alice's bonus first). We use the strategy method to elicit Bob's decisions. Instead of responding only when Alice asks for help, Bob makes a decision in advance, specifying whether he would help if asked. This ensures that Bob's choice is recorded regardless of Alice's actual decision. Alice is aware that Bob makes his decision in this way.



The choice of monetary parameters serves multiple objectives. First, we set the monetary cost of asking for help to be positive yet negligible to prevent extreme frequencies of asks (close to 0% or 100%). Second, the combination of asking and receiving help increases social surplus (from \$6.5 to \$7), while the combination of asking and not receiving help decreases it (from \$6.5 to \$6). We rely on such a setup to create additional incentives for helpers to help. Finally, if a helper chooses to help, they still earn more than their asker (\$4 vs. \$3), preserving altruistic incentives.

Round 2: Beliefs

In Round 2, Alice answers two questions in a random order:

- Question 1: 'What is the chance that Bob agrees to help?' **[*beliefs about help*]**
- Question 2: 'Think about all participants that were assigned to Group A. What is the chance that a randomly selected participant from Group A asks for help?' **[*beliefs about others*]**

The first question concerns the expectation of receiving help. If participants aim to maximize their expected payoff, this belief should strongly predict their decision to ask. The second belief pertains to the likelihood that another asker will request help. This measure helps determine whether askers are influenced by the actions of others or whether they project their own behavior onto others.

In Round 2, Alice can earn either \$0 or \$3. We use the BDM method to incentivize truthful belief reporting (Becker et al., 1964). The procedure is explained to participants at the outset, with periodic reminders provided throughout the experiment.

Figure 2.1: Differences in askers' instructions across the treatments

(a) Treatment ID

[*a participant is randomly assigned to Group A; helpers are in Group B*]

Each participant in Group B has a three-digit ID number.

Now one of the participants in group B will be randomly selected as YOUR COUNTERPART.

NEXT

Your counterpart is #100

NEXT

If you are selected to receive the bonus, your choices and your counterpart's choices will determine your payments.

Here is how it works. Exactly five participants from Group A will be assigned the same counterpart from Group B. Then, one out of five members of Group A who have the same counterpart from Group B will be selected to receive the bonus. Thus, you will have a 20% chance of being selected to receive a bonus, and then #100 makes the decision specifically for you.

[*actions and payments are explained*]

(b) Treatment No ID

[*a participant is randomly assigned to Group A; helpers are in Group B*]

Each participant in Group B has a three-digit ID number. If you are selected to receive the bonus, the computer will select your counterpart from Group B **after you complete the study**.

If you are selected to receive the bonus, your choices and your counterpart's choices will determine your payments. Here is how it works. One of five participants from Group A will be selected to receive the bonus and then matched to their counterpart from Group B. Thus, you will have a 20% chance of being selected for receiving a bonus, and any of the participants from Group B can affect your payment.

[*actions and payments are explained*]

Note: See notes in the footnote on p. 62.

2.2 Implementation

The experiment was approved by the Caltech IRB (IR23-1316) and pre-registered (AsPredicted #133951). We conducted the experiment on the Prolific online platform in June 2023. Given its short duration, it was combined with another experiment unrelated to this project.[7]

We recruited a total of 480 participants, equally split between two treatments: 200 askers and 40 helpers in Treatment ID, and 200 askers and 40 helpers in Treatment No ID. Random assignment between roles was achieved by running sequential sessions that were indistinguishable in title and description. Within each treatment, we first conducted the session for askers (200 participants) followed by the session for helpers (40 participants). Each participant could join only one session. The sample included participants aged 18 to 65, and was gender-balanced and restricted to individuals residing in the United States with a high approval rating on Prolific.

All participants received a fixed $1 participation payment for completing the experiment. In addition, the average earnings were $0.36 for askers and $3.86 for helpers.[8] The average completion time for askers was 3.39 minutes in Treatment ID and 2.88 minutes in Treatment No ID, with most of the difference stemming from the time spent on instructions and deciding whether to ask for help.

2.3 Baseline Hypothesis

Our main hypothesis, which we test with the goal of rejecting, is grounded in the discussion of the related literature. The existing evidence largely suggests that increasing a helper's identifiability decreases the likelihood of asking for help. One possible explanation is the fear of rejection, which may be heightened when the helper is identifiable. Askers might worry more about being rejected by a specific, identified person than by an anonymous or statistical helper. Bénabou et al. (2025) link this fear to the perceived value of the person making the request.

A second possibility could be that *optimism about receiving help* may be greater when the helper is statistical rather than identifiable. Prior research suggests that people are more willing to bet on uncertain future outcomes than on those that have

[7] Details of the other project are available upon request.

[8] Although helpers cannot get less than $4 in the asking game (see page 63), many of them got less than $4 due to being randomly assigned to be paid for Round 2, where they needed to submit their beliefs about askers and other helpers and could get either $0 or $3.

already been determined (Brun and Teigen, 1990; Rothbart and Snyder, 1970), and tend to take more risks when making predictions rather than postdictions (Strickland et al., 1966). Identifiability may partially resolve uncertainty before the request is made. In contrast, when the helper remains anonymous, uncertainty persists, potentially allowing people to remain more optimistic about their chances of receiving help.

At the same time, social preferences may influence the effect of identifiability in different directions. Given that individuals tend to favor ex-ante fairness in decision-making (Andreoni, Aydin, et al., 2020), the perceived fairness of asking depends on how participants interpret fairness in the given context. On one hand, placing the burden of helping on an anonymous or statistical helper may seem more justifiable than asking a specific individual, making people less likely to request help from an identified person. On the other hand, knowing the identity of a counterpart may highlight disparities in starting endowments due to luck, which could, in turn, increase the likelihood of asking. Thus, whether identifiability increases or decreases asking behavior may depend on whether participants focus more on avoiding the imposition of a burden on an identifiable individual or on rectifying perceived inequalities.

H_0 The rates of asking in Treatments ID and No ID are the same.

3 BASELINE RESULTS

Figure 2.2a presents the average frequency of asking by treatment. Providing ID numbers of helpers to askers increases the average asking rate from 67% to 76.5% ($p = 0.046$ in a two-sided proportion test).

Result 1. Asking is significantly more frequent in Treatment ID (with identified helpers).

This is the main result of the paper. We now investigate the mechanism behind this effect by analyzing askers' beliefs. Our analysis is simplified by the fact that the distributions of beliefs are similar across the two treatments. First, average beliefs remain statistically unchanged: askers estimate their chances of receiving help

(*beliefs about help*) and the likelihood that a random asker makes a request (*beliefs about others*) at approximately 42% and 65%, respectively, in both treatments.[9] Second, we narrowed the sample to those who asked for help and compared their beliefs between the treatments. Similarly, we compared the beliefs of those who did not ask between the treatments. This analysis also reveals no significant differences. Figures 2.2b and 2.2c show that average beliefs are nearly identical, and Figures B.1–B.3 in the Appendix confirm that the belief distributions do not noticeably differ.

We can expect that individuals strongly rely on their *beliefs about help*—that is, their perceived chance that the counterpart will agree to help—when deciding whether to ask. Moreover, the act of asking is likely to reinforce these beliefs: those who ask may overstate their beliefs about help, while those who do not ask may understate them. Overall, we expect that individuals who choose to ask hold higher *beliefs about help* than those who do not.

However, as Figure 2.2b shows, we observe this relationship in Treatment ID but not in Treatment No ID. More specifically, in Treatment ID, the decision to ask and *beliefs about help* are positively correlated, with a coefficient of 0.21 (p-value $<$ 0.01). In contrast, in Treatment No ID, the correlation is not statistically significant. This suggests that, in Treatment No ID, the decision to ask is primarily driven by factors other than the perceived chance of receiving help.

Result 2. Unlike in Treatment ID, in Treatment No ID, the decision to ask is not correlated with the perceived chance of receiving help (i. e., with the *beliefs about help*).

The absence of a relationship in Treatment No ID eliminates the possibility that the asking decision is primarily guided by the maximization of expected payoff, which is proportional to *beliefs about help* after asking and constant otherwise. Therefore, we cannot claim that introducing identification into *any* asking-for-help context would necessarily lead to an increase in the asking rate. Instead, in this specific context, relying more on expected payoff—rather than on other (as yet unclear) factors—led to more frequent asking.

We now turn to the question of which factors might override the influence of expected payoff on the decision to ask. Among several possibilities, we focus on the perceived

[9] Table B.1 in the Appendix reports summary statistics for the full sample and by treatment.

chance that other askers request help—*beliefs about others*. This factor is directly related to social norms and the potential stigma associated with asking for help.

Figure 2.2c shows that, in both Treatment ID and Treatment No ID, individuals who ask for help hold, on average, higher *beliefs about others* than those who do not. This pattern suggests that, in Treatment No ID, the perceived chance that others are also asking may be one of the factors that crowds out the role of expected payoff in the decision to ask.

However, based on correlation alone, we cannot conclude that beliefs causally affect the asking decision. It is also possible that individuals simply project their own decision onto others. To address this concern, we introduce an additional treatment designed to demonstrate that *beliefs about others* have a positive causal effect on the decision to ask. We describe this additional treatment at the beginning of the next section.

In Table 2.1, we explore the relationship between asking and beliefs in more detail using a correlation analysis. We control for gender, age, overconfidence,[10] risk aversion,[11] and performance on simple logic tasks.[12] Table 2.1 presents the results.

The analysis of the correlation coefficients suggests that, in Treatment No ID, the role of *beliefs about help* is substantially overshadowed by the influence of *beliefs about others*: although the two variables have approximately the same variance (see Table B.1 in the Appendix), the coefficient on beliefs about others is twice as large. In contrast, in Treatment ID, both types of beliefs are equally strongly associated with the asking decision.

We conclude this section by noting that our results are primarily driven by women in the sample. Their asking rates were 83% in Treatment ID and 68.47% in Treatment No ID, compared to 70% and 65.4% for men, respectively.[13] The treatment

[10]Following (Ortoleva and Snowberg, 2015), we constructed two measures of overconfidence. First, we asked participants a simple factual question about the year the wired telephone was invented. Then, to capture overprecision, we used two approaches: a qualitative measure based on self-reported confidence, and a quantitative measure based on the participant's stated probability that their response is within 25 years of the correct answer.

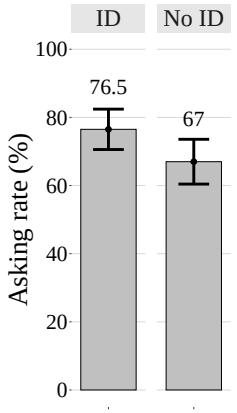
[11]We corrected for risk aversion measurement error using the Obviously Related Instrumental Variables (ORIV) method suggested by Gillen et al., 2019. In two separate tasks, participants were offered the opportunity to invest a portion of their 200 tokens in a risky project, with the gains and chances of success varying between tasks. We then used the two measures as instruments for each other to address measurement error in individual risk preferences.

[12]For example: "Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?"

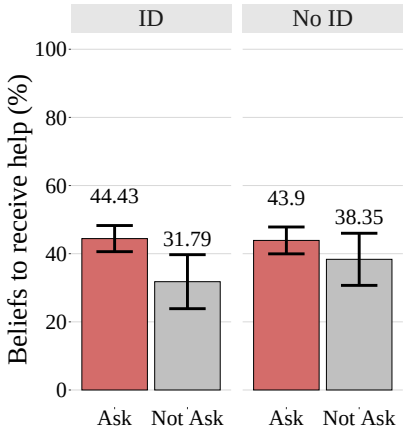
[13]Due to a technical issue on Prolific, we had to stop data collection for Treatment No ID

Figure 2.2: Average askers' responses: Baseline results

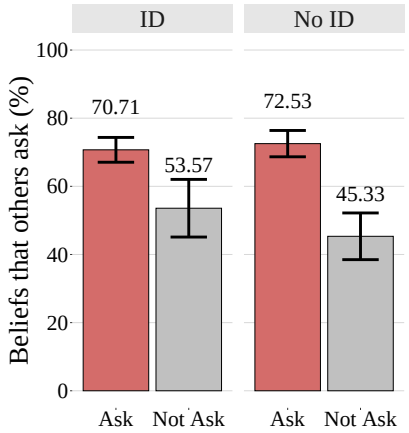
(a) Frequency of asks



(b) Beliefs about help



(c) Beliefs about others



Note: This figure shows the average responses to the primary experimental questions. The head of each subfigure gives the treatments. Treatment ID is for the askers with identified helpers, and Treatment No ID is for askers with non-identified helpers. In Figures 2.2b and 2.2c, the horizontal axis gives the subsamples within a treatment: those who ask for help and those who do not ask. Figure 2.2a gives the average frequencies of asks in both treatments; the rate of asking is 9.5 pp. higher in Treatment ID (76.5% vs. 67%; two-sided proportion test $p = 0.046$). Figure 2.2b gives the average perceived chance of receiving help. The difference between those who ask and do not is significant at 1% significance level within Treatment ID and not significant at 10% significance level within Treatment ID. The differences between the treatments within each decision are not significant at 10% level. Figure 2.2c gives the average perceived chance that a randomly selected asker asks for help. The difference within each treatment is significant at 1% level. The differences between the treatments within each decision are not significant at 10% level.

Table 2.1: Correlation between asking and beliefs

	Dependent variable: 1 if asks, 0 if does not ask							
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ID	0.095**	0.078**	0.023	−0.0004	0.279**	0.250***	0.290	0.218*
	(0.045)	(0.032)	(0.095)	(0.065)	(0.120)	(0.084)	(0.187)	(0.126)
Beliefs about help			0.002	0.001			0.006**	0.005***
			(0.001)	(0.001)			(0.002)	(0.002)
ID × Beliefs about help			0.002	0.002			−0.001	0.00001
			(0.002)	(0.001)			(0.004)	(0.003)
Beliefs about others					0.008***	0.008***	0.012***	0.011***
					(0.001)	(0.001)	(0.001)	(0.001)
ID × Beliefs about others					−0.003**	−0.003***	−0.006**	−0.005***
					(0.002)	(0.001)	(0.002)	(0.002)
Beliefs about help × Beliefs about others							−0.0001***	−0.0001***
							(0.00003)	(0.00003)
ID × Beliefs about help × Beliefs about others							0.0001	0.0001
							(0.0001)	(0.00004)
Observations	400	399	400	399	400	399	400	399
Controls	NO	YES	NO	YES	NO	YES	NO	YES
ORIV	NO	YES	NO	YES	NO	YES	NO	YES
Adjusted R^2	0.009		0.029		0.159		0.190	

Note: This table shows the correlation between asking and beliefs in the baseline treatment. ID is a binary variable equal to 1 if an individual is in Treatment ID and 0 if she is in Treatment No ID. Beliefs about help give the perceived chance of receiving help. Beliefs about others give the perceived chance that a randomly selected asker asks. Control variables include gender, age, overconfidence (Ortoleva and Snowberg, 2015), and performance in easy logic tasks. We also control for risk-aversion using ORIV method (Gillen et al., 2019). One of the participants in Treatment No ID retracted their demographic characteristics. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

difference persists even after controlling for risk aversion and other covariates. On the one hand, men exhibit a weaker relationship between their beliefs about others and their own decisions, which may explain the absence of further treatment effects among them: assigning ID numbers to helpers could not reduce an already weak connection. On the other hand, Croson and Gneezy (2009) show that women are generally more responsive to experimental interventions. Since assigning uninformative ID numbers is a subtle manipulation, we consider this sensitivity the most plausible explanation for the observed gender difference.

4 FOLLOW-UP EXPERIMENTS

This section describes the setup and results of the three follow-up experiments aimed at identifying the mechanism behind the ID effect on the frequency of asks. They were approved by Caltech IRB (IR23-1316A) and pre-registered (AsPredicted #144709[14] and #145442[15]). We ran the follow-up experiments on the Prolific online platform in September 2023. The criteria for the participants were the same as in the baseline study. In each of the experiments, the participants received a $1 fixed participation payment. Average earnings were $0.39 and $3.62 of askers and helpers, respectively.

4.1 Exogenous shift in *beliefs about others*

As Figure 2.2c and Table 2.1 show, *beliefs about others*—the perceived chance a randomly selected asker requests help—are strongly and positively associated with asking behavior in both treatments. However, the direction of causality between the two variables remains unclear. On the one hand, higher beliefs about the likelihood that others will ask may encourage individuals to ask themselves. On the other hand, individuals who decide to ask may project their own behavior onto others, thus generating the observed positive correlation. Since our primary interest lies in the decision-making process, we aim to determine whether higher *beliefs about others* causally increase the likelihood of asking.

and resume it in a new session. One of the 200 participants in Treatment No ID retracted their demographic information. In total, we had 100 women in Treatment ID and 92 women in Treatment No ID.

[14] https://aspredicted.org/RFV_PGB

[15] https://aspredicted.org/V31_SSW

To address this question, we exogenously increase *beliefs about others*. Specifically, in a follow-up experiment, we provide participants in both Treatments ID and No ID with the following information:

> We conducted the same experiment some time ago. More than 2/3 of the participants asked for help.

This statement is supported by the asking rates observed in the June sessions: 76.5% in baseline Treatment ID and 67% in baseline Treatment No ID. We recruited a total of 480 participants, with the same allocation across treatments as in the baseline experiment.

The intervention increased the average *beliefs about others* in Treatment No ID from 63.56% to 68.16%—a change that is statistically significant at the 5% level (two-sided t-test). In contrast, in Treatment ID, where the initial average was already close to $2/3$, we could not expect much further increase: beliefs rose slightly and insignificantly, from 66.69% to 67.64%.

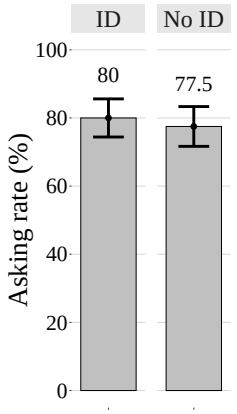
Figure 2.3 summarizes the average responses. The information intervention significantly increased the chance of asking in Treatment No ID (by 10.5 percentage points), eliminating the difference in asking rates between Treatments ID and No ID. In Treatment ID, the rate of asking increased by 3.5 percentage points after the intervention, but this change is not statistically significant. *Beliefs about help* did not change relative to the baseline results and exhibit the same patterns: they are predictive of the decision to ask in Treatment ID but do not differ between askers and non-askers in Treatment No ID. The results of the information intervention suggest that *beliefs about others* have a positive *causal* impact on one's own asking in Treatment No ID.

Result 3. *Beliefs about others*—that is, the perceived chance others ask—causally increase asking in Treatment No ID.

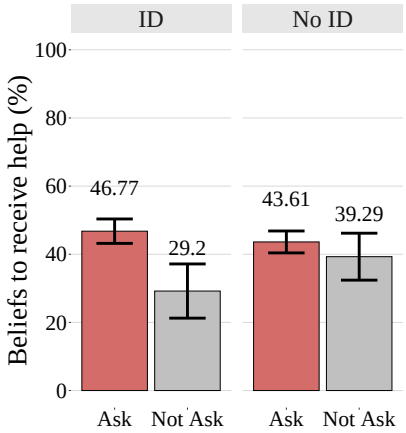
Results 1, 2, and 3 can be summarized as follows. When the helper has an ID, the expected payoff can predict the decision to ask. This is not the case when the helper does not have an ID: the influence of other factors dissolves the correlation between asking and the expected payoff altogether. One such factor is the belief about the

Figure 2.3: Average askers' responses after exogenous upward increase in *beliefs about others*

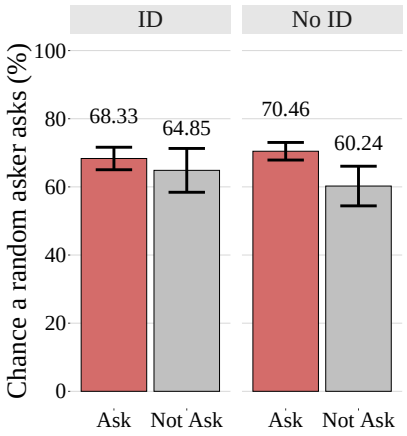
(a) Frequency of asks



(b) Beliefs about help



(c) Beliefs about others



Note: This figure shows the average responses to the primary experimental questions after they are told that more than two-thirds of the previous responders asked for help. The head of each subfigure gives the treatments. Treatment ID is for the askers with identified helpers, and Treatment No ID is for askers with non-identified helpers. In Figures 2.3b and 2.3c, the horizontal axes give the subsamples within a treatment: those who ask for help and those who do not ask. Figure 2.3a gives the average frequencies of asks in both treatments; the difference between the frequencies is insignificant. Figure 2.3b gives the average beliefs about help. The difference between those who ask and do not is significant at 1% level within Treatment ID and not significant at 10% level within Treatment ID. The difference between the treatments among those who do not ask has a p-value of 0.06, and it is insignificant at 10% level among those who ask. Figure 2.3c gives the average perceived chance that a randomly selected asker asks. The difference is insignificant at 10% level in Treatment ID and significant at 1% level in Treatment No ID. Each treatment is significant at 1% level. The differences between the treatments within each decision are not significant at 10% level.

likelihood that others ask for help—it has a positive and significant effect on an individual's decision to ask.

4.2 Ensuring askers feel anonymous to helpers

We designed the experiment so that any differences between treatments would stem solely from the askers' side. Failing to achieve this would risk confounding the interpretation of treatment effects. For instance, we would have to consider interpretations of *beliefs about others* that go beyond social norms. In Treatment No ID, it is clear that the helper's decision depends on the probability that a random asker requests help. In contrast, in Treatment ID, if an asker believes that she is identifiable to her helper, she may think that the helper's decision depends on the probability that *this particular* asker requests help.[16] If askers in Treatment ID do not believe they are treated as random individuals, this could offer an alternative explanation for why *beliefs about others* are less correlated with asking behavior in Treatment ID than in Treatment No ID (Table 2.1).

Therefore, we conducted a follow-up experiment to ensure that askers correctly understood the identification mechanism in the original design. We recruited 200 participants to repeat the baseline Treatment ID, with the following clarification added to the instructions: "[Your helper] does not know the ID of the [asker] they will be matched with!" We found no difference between the baseline Treatment ID and the new data. Thus, we conclude that askers felt anonymous to their helpers.

In addition, we explored whether askers in Treatment ID distinguished between their own chance of receiving help and the chance that other askers would receive help. We ran an experiment identical to the baseline Treatment ID, except that we added a third belief question in Round 2: "Think about all participants that were assigned to Group B, not only [your helper]. What is the chance that a randomly selected participant from Group B agrees to help?" We found that participants who asked for help were slightly but significantly more optimistic about their own helper compared to others. We further ran one more session in which decision-making and belief reports were submitted simultaneously. In this latest session, we found no difference between participants' beliefs about receiving help

[16] Analogously, a firm's production decisions differ when operating in a competitive market versus as a monopolist.

themselves and the help others would receive. We interpret the earlier difference as an artifact of having committed to the asking decision before stating beliefs.

This allows us to attribute the observed effects of identification to changes in askers' behavior that are unrelated to any changes in their perceptions of how helpers make decisions.

5 DISCUSSION AND CONCLUSION

This paper proposes an experiment to examine how a helper's identifiability affects the likelihood of receiving a request for help. The form of identifiability we study is extremely weak: the asker is merely provided with an uninformative ID number of the helper. Yet, the observed effect is substantial both economically and statistically—when the helper is identified, the probability of an ask increases by 9.5 percentage points.

We also elicit two types of beliefs from receivers: (1) their belief about the likelihood of receiving help (*beliefs about help*), and (2) their belief about the likelihood that a random receiver would ask (*beliefs about others*). While the chance of asking should, intuitively, be strongly correlated with the *beliefs about help*, we find this relationship only when askers have their helpers identified. If the helpers are statistical, the relationship between asking and *beliefs about help* is completely dissolved by other factors to the extent that they are not correlated. We find that *beliefs about others* are part these factors: they positively affect the chance of asking while the helpers are statistical.

The reasons suggested to drive the identifiable victim effect might offer insight into why the dynamic between decisions and beliefs shifts after ID assignments. Given the inherently social nature of *beliefs about others*, our findings align with the narrative proposed by Small and Loewenstein, 2003, where identification reduces social distance. This reduction in social distance is also closely tied to the well-documented emotional divergence in responses to identifiable versus statistical victims (S. Lee and Feeley, 2018; Small, 2015, August). Since our results highlight the centrality of *beliefs about others*, we interpret the weakening of the effect as a consequence of removing ID numbers, which increases the perceived social distance between the asker and her potential helper. Greater social distance from the helper, in turn, amplifies the salience of other askers' behavior, making it a more influential benchmark.

We conclude that providing even minimal identification of helpers allows askers to abstract from concerns unrelated to their expected payoff—concerns that might otherwise discourage them from seeking help. In our setting, this led to a higher frequency of asking. We expect similar effects in environments where there is a strong stigma around asking for help, despite the potential benefits to recipients. Fundraising platforms like GoFundMe may fit this profile. Their performance could potentially be enhanced by revealing limited, non-sensitive information about donors. Help-seeking might also be promoted in schools and workplaces by emphasizing the specific people who can be asked. Given that our intervention involved only a very weak form of identifiability—uninformative ID numbers—our findings suggest that such positive effects can be achieved without compromising the privacy of helpers.

References

Aimone, J. A., & Houser, D. (2012). What you don't know won't hurt you: A laboratory analysis of betrayal aversion. *Experimental Economics*, *15*(4), 571–588. https://doi.org/10.1007/s10683-012-9314-z (cit. on pp. 59, 61).

Andreoni, J., Aydin, D., Barton, B., Bernheim, B. D., & Naecker, J. (2020). When Fair Isn't Fair: Understanding Choice Reversals Involving Social Preferences. *Journal of Political Economy*, *128*(5), 1673–1711. https://doi.org/10.1086/705549 (cit. on pp. 61, 67).

Andreoni, J., & Rao, J. M. (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics*, *95*(7-8), 513–520. https://doi.org/10.1016/j.jpubeco.2010.12.008 (cit. on p. 58).

Becker, G. M., Degroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232. https://doi.org/10.1002/bs.3830090304 (cit. on p. 64).

Bénabou, R., Jaroszewicz, A., & Loewenstein, G. (2025). It hurts to ask. *European Economic Review*, *171*, 104911. https://doi.org/10.1016/j.euroecorev.2024.104911 (cit. on pp. 59–61, 66).

Bhargava, S., & Manoli, D. (2015). Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review*, *105*(11), 3489–3529. https://doi.org/10.1257/aer.20121493 (cit. on p. 60).

Bohns, V. K. (2016). (Mis)Understanding Our Influence Over Others: A Review of the Underestimation-of-Compliance Effect. *Current Directions in Psycho-*

logical Science, *25*(2), 119–123. https://doi.org/10.1177/0963721415628011 (cit. on p. 60).

Bohns, V. K., & Flynn, F. J. (2010). "Why didn't you just ask?" Underestimating the discomfort of help-seeking. *Journal of Experimental Social Psychology*, *46*(2), 402–409. https://doi.org/10.1016/j.jesp.2009.12.015 (cit. on p. 60).

Brun, W., & Teigen, K. H. (1990). Prediction and postdiction preferences in guessing. *Journal of Behavioral Decision Making*, *3*(1), 17–28. https://doi.org/10.1002/bdm.3960030103 (cit. on pp. 61, 67).

Chandrasekhar, A., Golub, B., & Yang, H. (2018, October). *Signaling, Shame, and Silence in Social Learning* (tech. rep. No. w25169). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w25169 (cit. on p. 60).

Chen, Y., & Gao, L. (2022). The Identified Donor Effect: Disclosure of the Donor's Name Shapes the Recipient's Behavior. *Journal of Consumer Psychology*, *32*(2), 232–250. https://doi.org/10.1002/jcpy.1243 (cit. on pp. 58–60).

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–74. https://doi.org/10.1257/jel.47.2.448 (cit. on p. 72).

Genevsky, A., Västfjäll, D., Slovic, P., & Knutson, B. (2013). Neural Underpinnings of the Identifiable Victim Effect: Affect Shifts Preferences for Giving. *The Journal of Neuroscience*, *33*(43), 17188–17196. https://doi.org/10.1523/JNEUROSCI.2348-13.2013 (cit. on p. 59).

Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. *Journal of Political Economy*, *127*(4), 1826–1863. https://doi.org/10.1086/701681 (cit. on pp. 69, 71).

Jaroszewicz, A., Loewenstein, G., Canavan, K., Martin, J., & Tevar, A. (2022). *The Psychological Pain of Asking for Live Kidney Donations* (Working paper). (Cit. on pp. 58–60).

Jenni, K., & Loewenstein, G. (1997). Explaining the Identifiable Victim Effect. *Journal of Risk and Uncertainty*, *14*(3), 235–257. https://doi.org/10.1023/A:1007740225484 (cit. on pp. 59, 61).

Lee, F. (1997). When the Going Gets Tough, Do the Tough Ask for Help? Help Seeking and Power Motivation in Organizations. *Organizational Behavior and Human Decision Processes*, *72*(3), 336–363. https://doi.org/10.1006/obhd.1997.2746 (cit. on pp. 58, 60).

Lee, S., & Feeley, T. H. (2016). The identifiable victim effect: A meta-analytic review. *Social Influence*, *11*(3), 199–215. https://doi.org/10.1080/15534510.2016.1216891 (cit. on p. 61).

Lee, S., & Feeley, T. H. (2018). The Identifiable Victim Effect: Using an Experimental-Causal-Chain Design to Test for Mediation. *Current Psychology*, *37*(4), 875–885. https://doi.org/10.1007/s12144-017-9570-3 (cit. on pp. 59, 61, 76).

Nadler, A. (2015). The Other Side of Helping. In D. A. Schroeder & W. G. Graziano (Eds.), *The Oxford Handbook of Prosocial Behavior*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.004 (cit. on p. 59).

Ortoleva, P., & Snowberg, E. (2015). Overconfidence in Political Behavior. *American Economic Review*, *105*(2), 504–535. https://doi.org/10.1257/aer.20130921 (cit. on pp. 69, 71).

Rothbart, M., & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, *2*(1), 38–43. https://doi.org/10.1037/h0082709 (cit. on pp. 61, 67).

Ryan, A. M., Gheen, M. H., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' social–emotional role, and the classroom goal structure. *Journal of Educational Psychology*, *90*(3), 528–535. https://doi.org/10.1037/0022-0663.90.3.528 (cit. on pp. 58, 60).

Small, D. A. (2015, August). On the Psychology of the Identifiable Victim Effect. In I. G. Cohen, N. Daniels, & N. Eyal (Eds.), *Identified versus Statistical Lives* (pp. 13–23). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217471.003.0002 (cit. on pp. 61, 76).

Small, D. A., & Loewenstein, G. (2003). Helping a Victim or Helping the Victim: Altruism and Identifiability. *Journal of Risk and Uncertainty*, *26*(1), 5–16. https://doi.org/10.1023/A:1022299422219 (cit. on pp. 58, 59, 61, 76, 82).

Song, J. (, Huang, J. (, & Jiang, Y. (2022). Mitigating the negative effects of service failure through customer identification. *Psychology & Marketing*, *39*(4), 715–725. https://doi.org/10.1002/mar.21615 (cit. on p. 59).

Strickland, L. H., Lewicki, R. J., & Katz, A. M. (1966). Temporal orientation and perceived control as determinants of risk-taking. *Journal of Experimental Social Psychology*, *2*(2), 143–151. https://doi.org/10.1016/0022-1031(66)90075-8 (cit. on pp. 61, 67).

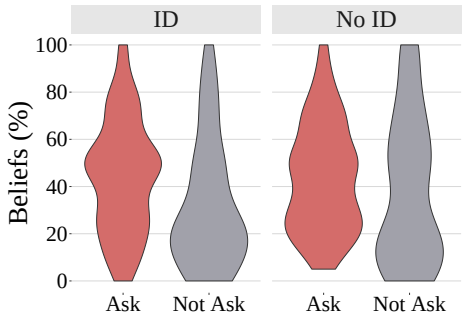
Zhao, H., Xu, Y., Li, L., Liu, J., & Cui, F. (2024). The neural mechanisms of identifiable victim effect in prosocial decision-making. *Human Brain Mapping*, *45*(2), e26609. https://doi.org/10.1002/hbm.26609 (cit. on p. 59).

B APPENDIX FOR CHAPTER 2

B.1 Askers in the Baseline Experiment

Figure B.1: Distributions of beliefs

(a) Beliefs about receiving help



(b) Beliefs about others to ask

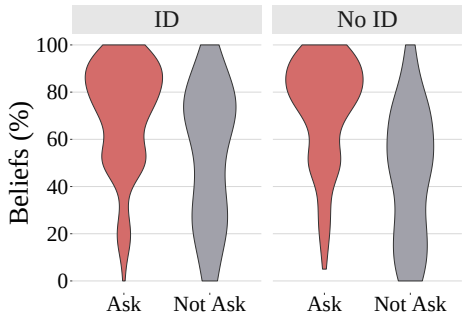
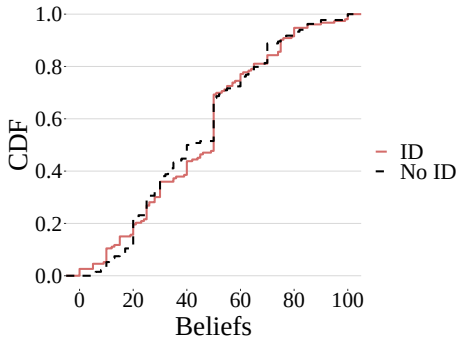


Figure B.2: CDFs of askers' beliefs about receiving help

(a) Ask



(b) Not ask

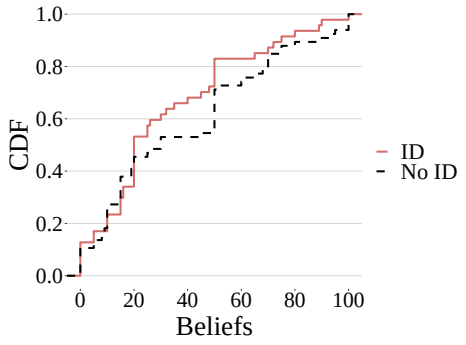
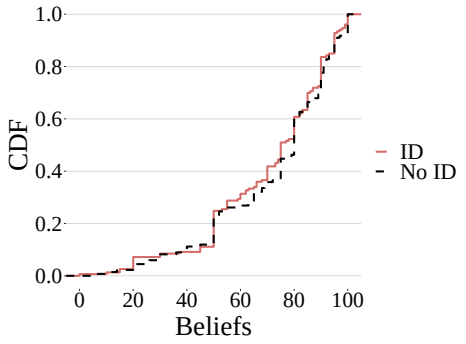


Figure B.3: CDFs of askers' beliefs about others to ask

(a) Ask



(b) Not ask

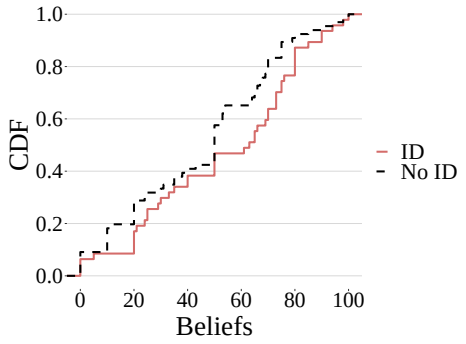


Table B.1: Askers' responses by treatment and decision

	Treatment ID							Treatment No ID						
	N	Mean	SD	Median	Min	Max	SE	N	Mean	SD	Median	Min	Max	SE
Share of those who decide to ask (%)														
Decide to ask	200	76.50	42.51	100	0	100	3.01	200	67.00	47.14	100	0	100	3.33
Belief about receiving help (%)														
All	200	41.46	25.25	43	0	100	1.79	200	42.07	26.05	40	0	100	1.84
Who ask	153	44.43	24.01	50	0	100	1.94	134	43.90	23.07	40.5	5	100	1.99
Who don't	47	31.79	26.99	20	0	100	3.94	66	38.35	31.11	30	0	100	3.83
Beliefs about others to ask (%)														
All	200	66.69	25.34	74	0	100	1.79	200	63.56	27.56	70	0	100	1.95
Who ask	153	70.71	22.80	75	0	100	1.84	134	72.53	22.61	80	5	100	1.95
Who don't	47	53.57	28.81	63	0	100	4.20	66	45.33	27.87	50	0	100	3.43

B.2 Ensuring askers feel anonymous to helpers: Data

One of the follow-up experiment explores the possibility that askers misunderstood the identification. Small and Loewenstein, 2003 show that assigning an uninformative ID to the victim in the dictator game increases giving. If the askers in my study misunderstood that only helpers are identified, the ID effect on asking could be driven by how askers perceive helpers.

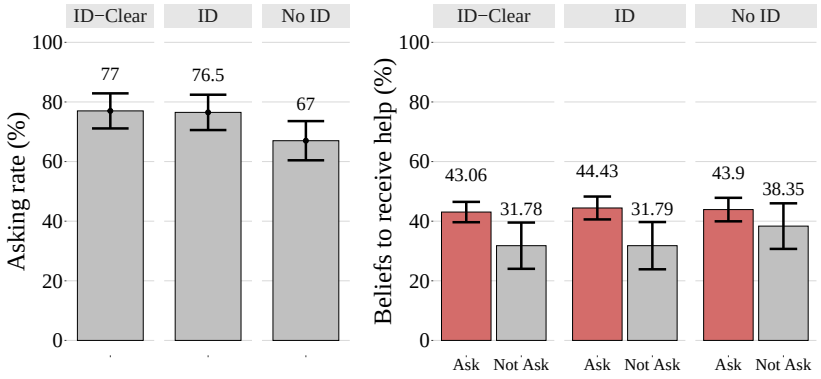
We recruited 200 participants to repeat the baseline Treatment ID with the addition of the following statement:

> [Your helper] does not know the ID of the [asker] they will be matched with!

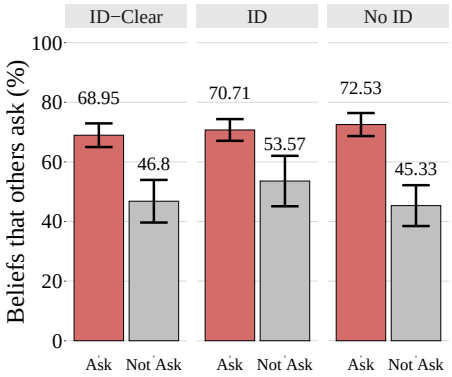
Figure B.4 reports the results of this treatment, which are not statistically different from the baseline Treatment ID. Further comparison of distributions of responses also did not reveal statistically significant differences. Therefore, askers did not perceive helpers differently across Treatments ID and No ID.

Figure B.4: Average askers' responses: Clarification treatment

(a) Frequency of asks (b) Beliefs about help



(c) Beliefs about others



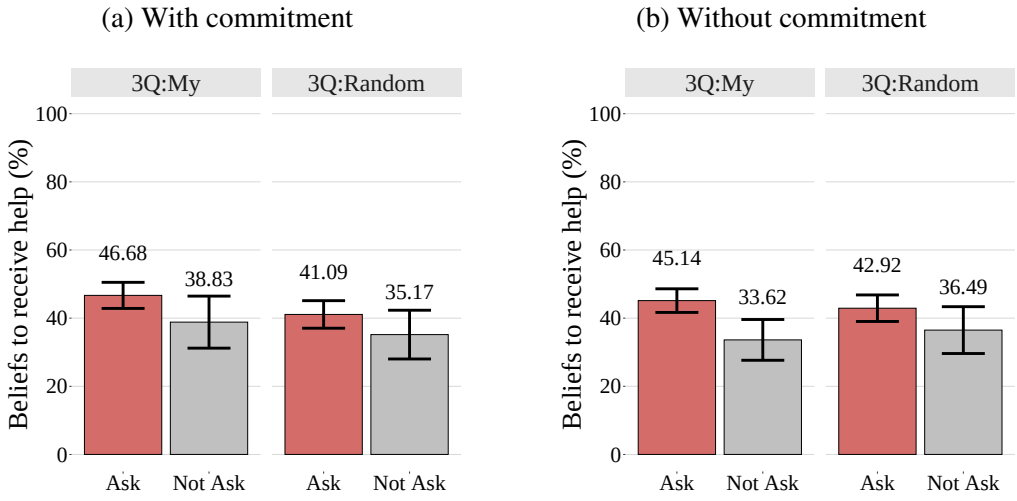
Note: This figure shows the average responses in the follow-up treatment for askers with ID, who receive clarification that their helper does not know their ID. These averages are denoted as 'ID-Clear', and they are depicted along the average responses of askers for the baseline treatments ('ID' and 'No ID'). The average responses for the baseline treatments are the same as in Figure 2.2 and are given for the comparison only. The head of each subfigure gives the treatments. In Figures 2.2b and 2.2c, the horizontal axes give the subsamples within a treatment: those who ask for help and those who do not ask. Across all three primary questions, the average responses of the askers in Treatment ID-Clear are not significantly different from the responses in Treatment ID of the baseline (at 10% level). It suggests that the clarification made no effect.

In another follow-up experiment, we first ran a treatment identically to baseline Treatment ID except for the new additional question in Round 2:

> Think about all participants that were assigned to Group B, not only [your helper]. What is the chance that a randomly selected participant from Group B agrees to help?

We call this session 'Beliefs with commitment', because the participants had to submit their decision about asking for help before reporting their beliefs. We found that the participants who ask for help are significantly more optimistic about their own helper than about others (Figure B.5). We then ran the 'Beliefs without commitment' session: all experimental questions were on one page. Without commitment, askers did not believe their helpers's giving to be different from others.

Figure B.5: Average beliefs about help: My helper vs. Random helper



Note: This figure shows the average perceived chance of receiving help in the follow-up treatment for askers with ID. They receive an additional question about the chance a random helper—not necessarily the one paired with them—agrees to help. Figure B.5a gives the average beliefs in the session in which the askers could not change their decision to ask or not after they saw the questions about beliefs. Figure B.5b gives the average beliefs in the session in which the askers could not change their decision. The head of each subfigure gives the question: '3Q:My' is about the helper that the asker is paired with, and '3Q:Other' is about a random helper. The horizontal axes give the subsamples within a treatment: those who ask for help and those who do not ask. In the session with commitment to the decision to ask or not, the difference between the questions is significant at 5% level among those who ask for help. This difference is not significant at 10% level among those who do not ask. At the same time, in the session without commitment, there are is no significant difference in the average beliefs between the questions within either of these subsamples. The difference between samples for the question about one's own helper is similar to that in the original sample (see Figure 2.2b): the beliefs are larger among those who ask (with p-value 0.07 with commitment and at 1% without commitment). The difference between samples for the question about a random helper is not significant at 10% either with or without commitment.

B.3 Helpers

In total, I ran seven identical sessions for helpers: two for the baseline treatment in June 2023 (80 observations in total) and five for the follow-up treatments in September 2023 (200 observations in total). I combined the data within each series of sessions to show the main characteristics of the responses.

In the baseline experiment ('Jun'), the average rate of help was 65%, and in the follow-up treatments ('Sep') it was 75.5%. The difference is insignificant at 5% significance level (p-value of a two-sided t-test is 0.091). This difference can be attributed to chance, especially considering that two out of five sessions in September had helping rates similar to those in June: 62.5% and 67.5%.

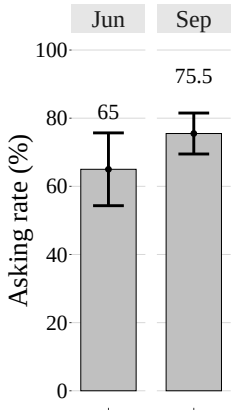
The helpers submitted two kinds of beliefs:

- 'Beliefs about being asked': 'What is the chance that your counterpart will ask you for help?'
- 'Beliefs about others to help': 'Think about all participants that were assigned to Group B. What is the chance that a randomly selected participant from Group B agreed to help?'

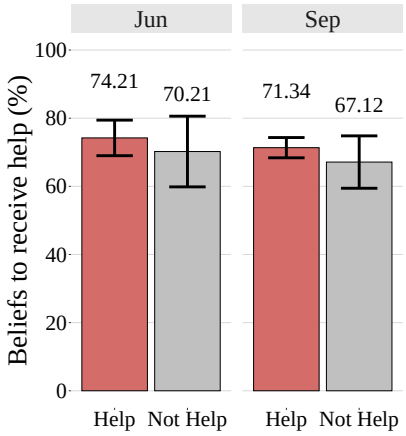
There is no systematic distinction in the distribution of beliefs between the two series, which also supports that the difference in the frequencies of help is due to chance.

Figure B.6: Average helpers' responses

(a) Frequency of help



(b) Beliefs about being asked



(c) Beliefs about others to help

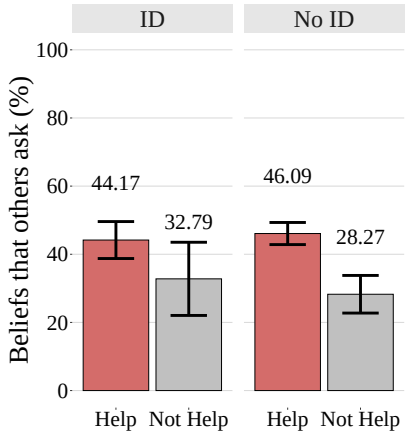
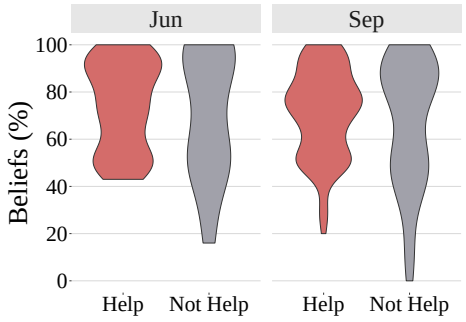


Figure B.7: Distributions of beliefs

(a) Beliefs about being asked



(b) Beliefs about others to help

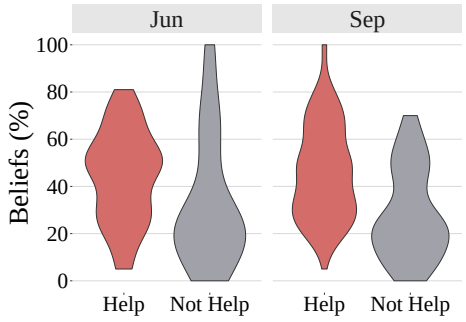
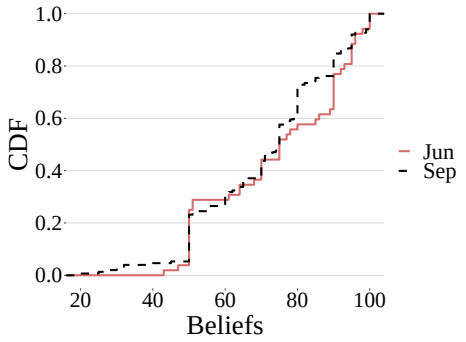


Figure B.8: CDFs of helpers' beliefs about being asked

(a) Among who helps



(b) Among who does not help

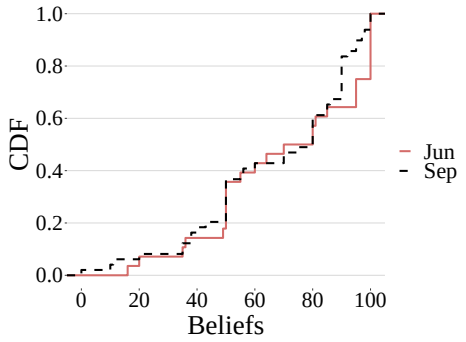
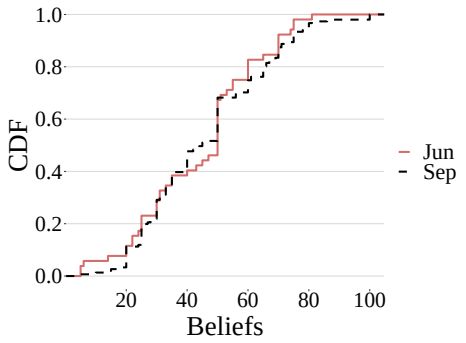
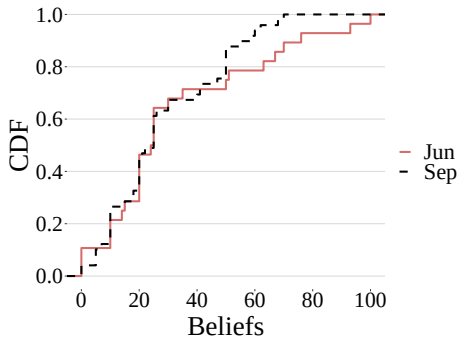


Figure B.9: CDFs of helpers' beliefs about others to help

(a) Among who helps



(b) Among who does not help



B.4 Instructions of the Baseline Experiment

The instructions begin with a common screen for all participants — **STRUCTURE** — that descrives the two groups of participants: Group A and Group B. After this screen, the instructions are different for three categories of participants: Group A in Treatment I, Group A in Treatment S, and Group B.

This experiment was part of a larger experiment, and it was Part IV in it. I removed 'Part IV' from the screen titles for clarity. At the beginning of the experiment, the participants were also explained why they should have submitted their best assessment when asked about their beliefs. Specifically, they were demonstrated how the binarized scoring rule works. I do not include this part in the instructions.

0. STRUCTURE

This part has 2 rounds. After completing this part of the experiment, you will receive **$1** in addition to your participation fee. You can also earn a few more dollars as a bonus. We will refer to it as a **potential bonus**.

You will be randomly assigned to **Group A** or **Group B**. There will be five times more participants in **Group A** than in **Group B**. Each participant from **Group A** will begin this part with a potential bonus of **$0.5**, and each participant from Group B will begin with a potential bonus of **$6**.

CHOOSE YOUR GROUP RANDOMLY

You are in **Group X**

GROUP A TREATMENT I

A-ID.1. Your bonus payment for this part

You are in **Group A**. You will have a 20% chance of being selected to receive a bonus. If you are selected to receive the bonus, you will be paid your potential bonus exactly for one of the two rounds, randomly determined. If Round 2 is selected for a bonus, you will be paid for one randomly selected question out of the two.

Each participant in Group B has a three-digit ID number.

Now one of the participants in group B will be randomly selected as YOUR COUNTERPART.

NEXT

Your counterpart is #100

NEXT

If you are selected to receive the bonus, your choices and your counterpart's choices will determine your payments. Here is how it works. Exactly five participants from Group A will be assigned the same counterpart from Group B. Then, one out of five members of Group A who have the same counterpart from Group B will be selected to receive the bonus. Thus, you will have a 20% chance of being selected to receive a bonus, and then #100 makes the decision specifically for you.

It is in your best interest to answer every question carefully and in a way that reflects what you truly prefer because these answers may determine your payment in this Part of the experiment.

A-ID.2. Round 1

YOUR INSTRUCTIONS

You begin the experiment with a potential bonus of **$0.5**. #100 began the experiment with a potential bonus of **$0.5**.

Remember that, among all participants in Group B, only #100 can affect your payment! You will ALWAYS learn if #100 decides to help you – we will send you a message after you complete the study.

You now have an opportunity to ask #100 for help. If you do not ask, your potential bonus will remain **$0.5** and #100's potential bonus will remain **$6**.

If you ask #100 for help, #100 will decide on both potential bonuses. If #100 agrees to help, your potential bonus will be **$3**, and their potential bonus will be **$4**. If #100 refuses to help, your potential bonus will be **$0**, and their potential bonus will be **$6**.

Click NEXT to see the instructions that all participants from Group B, including #100, see.

NEXT

INSTRUCTIONS FOR GROUP B

You begin this experiment with a potential bonus of **$6**. Your counterpart begins the experiment with a potential bonus of **$0.5**.

Your counterpart can ask you for help or not ask:

> If your counterpart does not ask you for help, you keep your potential bonus of **$6**, no matter what you choose. They will also keep their potential bonus of **$0.5**.
>
> If your counterpart asks you for help, you get to decide on both bonuses. If you help, you will get a potential bonus of **$4**, and they will get a potential bonus of **$3**. If you do not help, you will get a potential bonus of **$6**, and they will get a potential bonus of **$0**.

Question:

Will you ask #100 for help?

- Yes, I will ask.
- No, I will not ask.

NEXT

A-No-ID.3. Round 2

In this round, you will evaluate some chances. If you are selected to receive a bonus, and Round 2 is chosen for the bonus payment, you may receive **$3** depending on your answers below. Remember, the questions are designed so that your highest chance of receiving **$3** is achieved when you state **your best assessment** in each question.

Below we **repeat** the instructions from the previous round. Click NEXT to proceed to the questions.

YOUR INSTRUCTIONS

You begin the experiment with a potential bonus of **$0.5**. #100 began the experiment with a potential bonus of **$0.5**.

Remember that, among all participants in Group B, only #100 can affect your payment! You will ALWAYS learn if #100 decides to help you – we will send you a message after you complete the study.

You now have an opportunity to ask #100 for help. If you do not ask, your potential bonus will remain **$0.5** and #100's potential bonus will remain **$6**.

If you ask #100 for help, #100 will decide on both potential bonuses. If #100 agrees to help, your potential bonus will be **$3**, and their potential bonus will be **$4**. If #100 refuses to help, your potential bonus will be **$0**, and their potential bonus will be **$6**.

Click NEXT to see the instructions that all participants from Group B, including #100, see.

NEXT

INSTRUCTIONS FOR GROUP B

You begin this experiment with a potential bonus of **$6**. Your counterpart begins the experiment with a potential bonus of **$0.5**.

Your counterpart can ask you for help or not ask:

> If your counterpart does not ask you for help, you keep your potential bonus of **$6**, no matter what you choose. They will also keep their potential bonus of **$0.5**.
>
> If your counterpart asks you for help, you get to decide on both bonuses. If you help, you will get a potential bonus of **$4**, and they will get a potential bonus of **$3**. If you do not help, you will get a potential bonus of **$6**, and they will get a potential bonus of **$0**.

Question 1: What is the chance that #100 agrees to help?

————————- % SLIDER ————————-

Question 2: Think about all participants that were assigned to Group A. What is the chance that a randomly selected participant from Group A asked for help?

————————- % SLIDER ————————-

NEXT

GROUP A TREATMENT S

A-No-ID.1. Your bonus payment for this part

You are in **Group A**. You will have a 20% chance of being selected to receive a bonus. If you are selected to receive the bonus, you will be paid your potential bonus exactly for one of the two rounds, randomly determined. If Round 2 is selected for a bonus, you will be paid for one randomly selected question out of the two.

Each participant in Group B has a three-digit ID number. If you are selected to receive the bonus, the computer will select your counterpart from Group B **after you complete the study**.

If you are selected to receive the bonus, your choices and your counterpart's choices will determine your payments. Here is how it works. One of five participants from Group A will be selected to receive the bonus and then matched to their counterpart from Group B. Thus, you will have a 20% chance of being selected for receiving a bonus, and any of the participants from Group B can affect your payment.

It is in your best interest to answer every question carefully and in a way that reflects what you truly prefer because these answers may determine your payment in this Part of the experiment.

A-No-ID.2. Round 1

YOUR INSTRUCTIONS

You begin the experiment with a potential bonus of **$0.5**. All participants in Group B began the experiment with a potential bonus of **$6**.

Remember that **any of the participants in Group B can affect your payment!** You will ALWAYS learn if a random participant from Group B decides to help — we will send you a message after you complete the study.

You now have an opportunity to ask for help. If you do not ask, your potential bonus will remain **$0.5**. After you complete the study, the computer will randomly choose a participant in Group B for their potential bonus to remain **$6**.

If you ask for help, after you complete the study, the computer will randomly choose a counterpart for you. The decision of the counterpart will determine potential bonuses for you both. If they agree to help, your potential bonus will be **$3**, and their potential bonus will be **$4**. If they refuse to help, you will get a potential bonus of **$0**, and their potential bonus will remain **$6**.

Click NEXT to see the instructions which any participant from Group B see.

NEXT

INSTRUCTIONS FOR GROUP B

You begin this experiment with a potential bonus of **$6**. Your counterpart begins the experiment with a potential bonus of **$0.5**.

Your counterpart can ask you for help or not ask:

> If your counterpart does not ask you for help, you keep your potential bonus of **$6**, no matter what you choose. They will also keep their potential bonus of **$0.5**.
>
> If your counterpart asks you for help, you get to decide on both bonuses. If you help, you will get a potential bonus of **$4**, and they will get a potential bonus of **$3**. If you do not help, you will get a potential bonus of **$6**, and they will get a potential bonus of **$0**.

Question:

Will you ask for help?

- Yes, I will ask.
- No, I will not ask.

NEXT

A-No-ID.3. Round 2

In this round, you will evaluate some chances. If you are selected to receive a bonus, and Round 2 is chosen for the bonus payment, you may receive **$3** depending on your answers below. Remember, the questions are designed so that your highest chance of receiving **$3** is achieved when you state **your best assessment** in each question.

Below we **repeat** the instructions from the previous round. Click NEXT to proceed to the questions.

YOUR INSTRUCTIONS

You begin the experiment with a potential bonus of **$0.5**. All participants in Group B began the experiment with a potential bonus of **$6**.

Remember that **any of the participants in Group B can affect your payment!** You will ALWAYS learn if a random participant from Group B decides to help — we will send you a message after you complete the study.

You now have an opportunity to ask for help. If you do not ask, your potential bonus will remain **$0.5**. After you complete the study, the computer will randomly choose a participant in Group B for their potential bonus to remain **$6**.

If you ask for help, after you complete the study, the computer will randomly choose a counterpart for you. The decision of the counterpart will determine potential bonuses for you both. If they agree to help, your potential bonus will be **$3**, and their potential bonus will be **$4**. If they refuse to help, you will get a potential bonus of **$0**, and their potential bonus will remain **$6**.

Below are the instructions that all participants from Group B see.

NEXT

INSTRUCTIONS FOR GROUP B

You begin this experiment with a potential bonus of **$6**. Your counterpart begins the experiment with a potential bonus of **$0.5**.

Your counterpart can ask you for help or not ask:

> If your counterpart does not ask you for help, you keep your potential bonus of **$6**, no matter what you choose. They will also keep their potential bonus of **$0.5**.
>
> If your counterpart asks you for help, you get to decide on both bonuses. If you help, you will get a potential bonus of **$4**, and they will get a potential bonus of **$3**. If you do not help, you will get a potential bonus of **$6**, and they will get a potential bonus of **$0**.

Question 1: What is the chance that your counterpart agrees to help?

————————- % SLIDER ————————-

Question 2: Think about all participants that were assigned to Group A. What is the chance that a randomly selected participant from Group A asked for help?

————————- % SLIDER ————————-

NEXT

GROUP B (BOTH TREATMENTS)

A-ID.1. Your bonus payment for this part

This part has two rounds: Round 1 has one question, and Round 2 has two questions.

You will be paid your potential bonus exactly for one of the two rounds, randomly determined. If Round 2 is selected for a bonus, you will be paid for one randomly selected question out of the two.

It is in your best interest to answer every question carefully and in a way that reflects what you truly prefer because these answers may determine your payment in this Part of the experiment.

B.2. Round 1

You will be randomly matched with a participant from Group A. We will call this participant **your counterpart**. You will have an opportunity to help your counterpart if they ask you. Before making their decision, your counterpart learns the information given to you.

YOUR INFORMATION

You begin this experiment with a potential bonus of **$6**. Your counterpart begins the experiment with a potential bonus of **$0.5**.

Your counterpart can ask you for help or not ask:

> If your counterpart does not ask you for help, you keep your potential bonus of **$6**, no matter what you choose. They will also keep their potential bonus of **$0.5**.
>
> If your counterpart asks you for help, you get to decide on both bonuses. If you help, you will get a potential bonus of **$4**, and they will get a potential bonus of **$3**. If you do not help, you will get a potential bonus of **$6**, and they will get a potential bonus of **$0**.

Click NEXT to make your choice.

NEXT

Question:

Will you help your counterpart if they ask you?

- Yes, I will help.
- No, I will not help.

B.3. Round 2

In this round, you will evaluate some chances. As before, the questions are designed so that your highest chance of getting the bonus is achieved when you state **your best assessment**! Your potential bonus for this round will be $3.

Below we **repeat** the instructions from the previous round. Click NEXT to proceed to the questions.

YOUR INFORMATION

You begin this experiment with a potential bonus of **$6**. Your counterpart begins the experiment with a potential bonus of **$0.5**.

Your counterpart can ask you for help or not ask:

> If your counterpart does not ask you for help, you keep your potential bonus of **$6**, no matter what you choose. They will also keep their potential bonus of **$0.5**.
>
> If your counterpart asks you for help, you get to decide on both bonuses. If you help, you will get a potential bonus of **$4**, and they will get a potential bonus of **$3**. If you do not help, you will get a potential bonus of **$6**, and they will get a potential bonus of **$0**.

Question 1: What is the chance that your counterpart will ask you for help?

————————- % SLIDER ————————-

Question 2: Think about all participants that were assigned to Group B. What is the chance that a randomly selected participant from Group B agreed to help?

————————- % SLIDER ————————-

NEXT

Chapter 3

BELIEFS OF OTHERS: AN EXPERIMENT

1 INTRODUCTION

Information plays a pivotal role in strategic settings. New evidence compels people to revise their beliefs and adjust their actions accordingly. As a result, people care not only about their own beliefs but also about the beliefs of others, which are shaped by the information others observe. This is true in games with asymmetric information (Spence, 1973), coordination games (Morris and Shin, 2002), social learning settings (Bikhchandani et al., 2024), and global games (Carlsson and van Damme, 1993) among others. In all these environments, one's strategic incentives depend on how one expects new evidence to affect other players' beliefs.[1]

Much is known about how people update their own beliefs in response to new evidence (we survey this literature in Section 1.1). At the same time, little is known about how people think others update their beliefs. This is the focus of our paper. We study how second-order beliefs respond to information, i.e., how people think others update their beliefs when those others encounter new evidence. As argued above, this is a necessary building block of many strategic interactions, which makes this question particularly suitable for experimental investigation.

We conduct a series of experiments and empirically document how people think others revise their beliefs and how that relates to people's own belief updating process. Our study examines participants' genuine, home-grown beliefs about various factual statements—some neutral and rooted in general knowledge, while others politically charged. To simplify the exposition, throughout the paper, we refer to two players, Anne and Bob. Anne is tasked with predicting Bob's beliefs. In all treatments, Anne knows Bob's prior and the accuracy of the information structure from which Bob receives his signals. However, in some treatments, Anne directly observes the signal realization Bob receives, while in others, she does not. We refer to the former scenario as *Bob's conditional posterior* and the latter as *Bob's expected*

[1] In equilibrium, people are expected to predict correctly others' revised beliefs and their actions.

posterior.[2,3]

Our experimental design is inspired by the recent work of Navin Kartik et al., 2021 and, more broadly, by the principles of Bayesian updating. According to this theory, Anne's beliefs about Bob's conditional posterior should be independent of her own prior beliefs; it should only depend on Bob's prior and signal precision. The predictions about Bob's expected posterior are more nuanced. If Anne and Bob share the same prior, Bob's expected posterior should be the same as Bob's and Anne's priors. This prediction is a fundamental property of Bayesian updating: beliefs are Martingale, meaning that new information cannot systematically alter beliefs in any direction. However, if Anne and Bob have different priors, Navin Kartik et al., 2021 show that any new information will shift Bob's expected posterior closer to Anne's prior, with the gap between the two narrowing as the signal becomes more accurate. This theoretical result is known as information validates prior (IVP).[4] Motivated by these theoretical predictions, our experiment features variation in the distance between Anne's and Bob's priors and information structures with different accuracies.

We find strong support for the Martingale property: regardless of signal precision, when Anne and Bob share the same prior, Anne believes that Bob's expected posterior will be equal to his prior. Regarding the IVP property, we find partial support. In line with the IVP, Anne believes that *any* new evidence will decrease the disagreement between them, shifting Bob's expected posterior closer to her prior. This is true for all statements, including the politically charged ones. However, more precise information structures only marginally enhance this effect. We do observe a significant difference in the effectiveness of a more precise information structure for neutral statements when Anne's initial beliefs differ greatly from Bob's and when

[2]The two scenarios capture distinct but common situations. Imagine a friend who tells you the piece of news he read today. You are trying to predict how his prior beliefs would change in light of this news. This is predicting someone's conditional posterior. Alternatively, imagine a situation in which you are advising your friend to watch some news channel tonight, and, beforehand, trying to predict how that would change his beliefs. This is predicting someone's average posterior.

[3]We use the terms information structure accuracy, signal precision, and information quality interchangeably.

[4]These results hold in settings that satisfy standard ordering assumptions: the priors must be likelihood-ratio ordered, and the signal structures from which Bob draws new evidence must satisfy the monotone likelihood-ratio property. When the state is binary, as in our experiment, these assumptions are nonrestrictive. Moreover, for the binary state, the IVP property is equivalent to the result obtained in Francetich and Kreps, 2014, according to which conditional on the event being true, the expected posterior is bigger than the prior. However, as discussed in Navin Kartik et al., 2021, neither of the two results (Navin Kartik et al., 2021 and Francetich and Kreps, 2014) nests each other in a more general setting beyond binary signals.

Anne holds relatively extreme priors herself. Otherwise, Anne expects Bob's beliefs to be fairly rigid and not responsive to the quality of information he samples from, diverging from what Bayesian theory predicts.

To understand why Anne's beliefs about Bob's expected posteriors are less responsive to the information quality than expected, we study three elements that jointly determine expected posteriors. The first element is how Anne updates her own beliefs. In line with previous literature, we find that Anne tends to underinfer both from her prior and from new evidence. This underinference results in less responsive (flatter) than Bayesian posteriors and it is stronger for politically charged statements. Moreover, we present a novel finding indicating that corner beliefs are not as rigid and degenerate, as previously thought; Anne is willing to revise these beliefs when confronted with contradictory evidence.

The second element concerns Anne's beliefs about Bob's conditional posteriors — how she expects Bob to update his prior when she observes the signal he receives. Our findings reveal that Anne tends to project her own belief-updating process onto Bob. She believes that Bob underinfers both from his prior and from new evidence. Notably, Anne thinks Bob underinfers from his prior *more than she does* when updating her own beliefs. For political statements, Anne expects that new information will have little effect on Bob's priors. Similar to her own corner priors, Anne thinks that Bob's corner priors can shift when faced with contradictory evidence. Overall, these patterns lead to Bob's conditional posteriors being flatter than Bayesian ones, showing less sensitivity to changes in Anne's prior and remaining relatively similar regardless of the quality of the information he consumes — this constitutes the first "flattening" effect.

The third element shaping Bob's expected posterior is the signal distribution, which determines how Bob's conditional posteriors are weighted in the expected value. Compared to the Bayesian benchmark, Anne expects the signal frequencies to be less responsive to both her own prior and the quality of the information Bob receives — this constitutes the second "flattening" effect.

The two "flattening" effects operate in the same direction, collectively making Bob's expected posteriors quite rigid and less responsive to the quality of information he consumes relative to what Bayesian theory predicts. This finding is important as it underscores the limited impact of information in altering perceptions of others' beliefs, and consequently influencing their behavior. Given that information plays a vital role in economic settings, particularly as a tool for policy interventions, this

result challenges the presumption of its efficacy in driving collective action.

In our analysis, we use a combination of reduced-form analysis and structural estimations.[5] Many of the results discussed above are evident from the raw data without the need for a behavioral model. However, the structural approach provides a parsimonious framework to capture observed patterns and allows us to conduct counterfactual exercises, which often require extrapolation beyond the parameters directly observed in the experimental data.

In the first structural exercise, we compare the magnitudes of two "flattening" effects and find that Bob's non-responsiveness to information quality relative to the Bayesian benchmark is primarily due to a lack of sensitivity in his conditional posteriors to the quality of the information, rather than a flattening in signal frequencies. In the second structural exercise, we assess the magnitude of Anne's mistakes in predicting Bob's expected posteriors, comparing them to the *actual average Bob's posteriors*.[6] By and large, the mistakes are quite small indicating that Anne is remarkably good at predicting Bob's average posteriors. The largest mistakes she makes pertain to situations in which her own and Bob's priors are extreme and are on different sides of the spectrum. In these cases, Anne overestimates how much information will shift Bob's opinions toward her own. This result stems from Anne predicting that Bob underinfers from his prior to a larger extent than he actually does. This highlights the challenge of predicting how others respond to new information, particularly when strong and polarized opinions are involved.

Our findings extend beyond the settings discussed at the beginning of the introduction and offer broader insights into societal polarization. A substantial body of research in Political Science and Economics has focused on the drivers of polarization in the United States, which has deepened over recent decades, and explored potential solutions to mitigate it (McCarthy, 2019; McCarthy et al., 2006). The mere abundance of news sources, many of which exhibit some degree of political bias, does not alleviate polarization (Azzimonti and Fernandes, 2023; DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017). Individuals tend to select infor-

[5] We explore several prominent models from the literature and find that the model proposed by Grether, 1980 offers the best fit to our data, achieving this with the fewest parameters compared to alternatives. Other models we estimate include the social exchange model of Yuksel and Oprea, 2022, Woodford, 2020's model of cognitive imprecision, and the base-rate neglect model (see chapter 6 in Benjamin, 2019 survey for the evidence and theoretical underpinning of this phenomenon). Notably, the social exchange model of Yuksel and Oprea, 2022 allows Anne to update her belief upon learning Bob's prior. This possibility, however, does not alter the main results of the paper.

[6] This is different from the previous analysis, in which we took Bayesian prediction as a benchmark against which to compare Anne's predictions of Bob's average opinions after obtaining new evidence.

mation sources aligned with their pre-existing beliefs (Garrett, 2009; Stroud, 2010), reinforcing their prior convictions when consuming such content, in line with the martingale property of beliefs. A natural question arises: what if individuals were exposed to news from opposing political perspectives, such as Democrats reading Republican-aligned news? According to the IVP property, each group will have an incentive to invest in informing the other, so that the other begins to share its perspective. However, our results challenge this approach, showing that while exposure to different viewpoints does shift expectations about others' beliefs, the change is modest, unaffected by the quality of the news source, and people generally recognize its limited potential to reduce polarization.

1.1 Connection to the Literature

Information design. As discussed in the introduction, our experiment builds on the findings of Navin Kartik et al., 2021, which contribute to the extensive theoretical literature on information design. While we do not aim to provide a comprehensive review of this literature, we want to highlight a few studies that explore strategic communication in environments with heterogeneous priors. For example, Hirsch, 2016 examines a model in which a principal and an agent share common goals but hold heterogeneous prior beliefs about which policy is most effective. This disagreement complicates the agent's motivation but can be alleviated through policy experimentation and observing outcomes. Alonso and Camara, 2016 considers a setting where the sender and receiver have differing prior beliefs, and the sender designs an experiment to persuade the receiver. The authors characterize the set of posterior belief distributions that can be induced by such experiments in setup with flexible information structure, i.e., no standard ordering assumptions, and identify necessary and sufficient conditions under which persuasion benefits the sender. Che and N. Kartik, 2009 investigate a game in which a decision-maker consults an adviser before making a decision. The adviser can exert costly effort to obtain a signal about the state and communicate this information to the decision-maker. Although both parties care about the state, their differing prior beliefs create a tension: these differences incentivize information acquisition but simultaneously lead to information loss through strategic communication. In all these papers, agents must form beliefs about how others, who may hold different priors, update their beliefs when new evidence arrives—a process which we investigate experimentally in our study.

First-order beliefs. In recent decades, we have learned a lot about how people update their beliefs upon encountering new evidence. Benjamin, 2019 provides an excellent and comprehensive review of empirical research from both Economics and Psychology, identifying consistent patterns and notable deviations from Bayesian theory. While some findings support Bayesian predictions, others highlight systematic discrepancies. Recent contributions to the field include Esponda et al., 2023, Augenblick et al., 2024, Ba et al., 2023, Gneezy et al., 2023, Enke and Graeber, 2023, and Marina Agranov and Reshidi, 2024 among others. By and large, this literature finds that while belief revisions generally follow the direction predicted by Bayesian theory, the magnitude of these revisions often deviates from the expected levels.

Most studies in this branch of literature employ neutral contexts and induce participants' priors to establish a controlled baseline for initial beliefs. A notable exception is the recent study by Thaler, 2024, which elicits participants' genuine beliefs on politically charged topics such as crime, climate change, gun control, and racial discrimination. This study finds that individuals distort new information in favor of their pre-existing views, consistent with motivated reasoning mechanisms. Its design elegantly differentiates this explanation from Bayesian updating motives. Like Thaler, 2024, we use genuine beliefs in our experiment but pursue a different research question focusing on how people think others revise their genuine beliefs upon receiving new information.

Higher-order beliefs. Our paper contributes to a growing experimental literature that studies higher-order beliefs and higher-order rationality. Most of this literature focuses on strategic settings: higher-order beliefs play an important role in these settings as they affect what actions players take.[7] For instance, Manski and Neri, 2013 elicit the subjects' first- and second-order beliefs in the Hide-and-Seek game and examine the coherence between these beliefs and actions. The results show remarkable consistency: observed choices are optimal given first-order beliefs in 89% of the time and in 75% of the time given second-order beliefs. P. J. Healy, 2024 elicits participants' preferences over game outcomes, their strategies, as well as first- and second-order beliefs in a series of classical games, including Prisoners'

[7] There are several excellent surveys of belief elicitation in experiments including Trevino and Schotter, 2014, Charness, Gneezy, and Rasocha, 2014, Schlag et al., 2015, and P. Healy and Leo, 2024. The survey of Trevino and Schotter, 2014 provides a detailed discussion of elicitation methods used to recover second-order beliefs.

Dilemma and the Centipede game. The data reveals heterogeneity in participants' preferences, which are not captured by game payoffs, but this heterogeneity only partially explains the gap between participants' beliefs and their own actions. Terri Kneeland, 2015 studies the Ring Games and demonstrates that over 70% of players are both rational and believe in others' rationality, though this decreases for higher-order beliefs. Friedenberg and T. Kneeland, 2024 extend this work to distinguish between players who have limited reasoning abilities and those who can reason iteratively but have limited belief in others' rationality and find that over 60% of participants engage in strategic reasoning beyond basic rationality. Calford and Chakraborty, 2023 show that the discrepancies in one's belief about an opponent and one's beliefs about others' beliefs about that opponent affect deviations from subgame perfection in a sequential social dilemma. Szkup and Trevino, 2020 infer how people think others update their beliefs in a coordination game with incomplete information, i.e., the global game.

Another class of games in which second-order beliefs are crucial is psychological games. In these games players' payoffs depend not only on material payoffs but also on the first-, second-, and possible higher-order beliefs about one's opponent. For instance, Dufwenberg and Gneezy, 2000 elicit players' first- and second-order beliefs in the Lost Wallet game, Charness and Dufwenberg, 2006 do so in the Trust Game, and M. Agranov et al., 2024 do so in an extended version of the sender-receiver game.

Some of the papers discussed above infer second-order beliefs from participants' actions and participants' beliefs about others' actions, while others elicit second-order beliefs directly. Our paper uses the latter approach and elicits second-order beliefs directly without relying on inference techniques. However, different from this literature, we deliberately focus on a non-strategic environment (no interdependence), which provides a clean free-from-strategic-considerations playfield to document how people think others revise their beliefs in response to new information.

The two closely related yet distinct studies are Evdokimov and Garfagnini, 2022 and Trujano-Ochoa, 2024. Evdokimov and Garfagnini, 2022 investigate higher-order beliefs in a three-player game where participants receive either private or public signals about the state. Player 1 reports his beliefs, Player 2 reports second-order beliefs, and Player 3 reports third-order beliefs. The authors find that belief updating is slower with private information, and higher-order learning often fails. In contrast, we test key Bayesian properties—specifically, the Martingale and IVP

properties—and focus on how second-order beliefs respond to new information. Similar to our motivation, Trujano-Ochoa, 2024 explores how people expect others to update their beliefs. His focus is on establishing to what extent people consider the biases of others when updating their beliefs and on information acquisition patterns.[8]

Finally, our paper relates to a small experimental literature that studies whether people anticipate the biases of others. The recent paper by Danz, Madarasz, and Wang, 2024 studies the relationship between the extent to which one projects her information onto others and the extent to which one anticipates but underestimates the projection of others onto her as predicted by the behavioral model of Madarasz, 2016. The results show strong support for the projection equilibrium model. Fedyk, 2024 finds that individuals exhibit substantial sophistication regarding the present bias of others, while being mostly naive about their own.

The rest of the paper is structured as follows. We present the conceptual framework in Section 2. Section 3 describes our experimental design and the experimental procedures. Section 4 presents reduced-form evidence on Martingale and IVP properties. Section 5 utilizes all conducted treatments to unpack the aggregate results presented in the previous section. Section 6 offers some conclusions.

2 Conceptual Framework

Consider a standard belief-updating task with a binary state $\omega \in \{0, 1\}$. There are two decision-makers, Anne and Bob, who may have the same or different priors about the state. We denote by a_0 and b_0 the prior of Anne and Bob, respectively. These priors indicate the probability that the state is $\omega = 1$ according to each of the two decision-makers. Bob receives a partially informative signal s and updates his beliefs about the state. We denote by b_s Bob's posterior belief after observing signal s and refer to it as *Bob's conditional posterior*. The signals are also binary and have accuracy θ. The accuracy of a signal indicates the likelihood that the signal matches the state conditional on the state, i.e., $\theta = \Pr[s = \omega|\omega]$.

Anne knows both Bob's prior and signal accuracy, and she is tasked with predicting Bob's average posterior after he receives a signal. The challenge arises because

[8]This work is still in progress. The preliminary draft we have access to suggests that, on average, people expect others to update in a similar manner to themselves. However, they show a significantly lower willingness to pay when others' strategies are implemented on their behalf. This latter finding is consistent with an expectation that others are more conservative in updating than they are themselves.

Anne does not observe the signal Bob receives; instead, she must weigh Bob's conditional posteriors according to the likelihood of the signals. We call this object *Bob's expected posterior*, and denote it by $\mathbb{E}[b]$. If Anne is Bayesian and expects Bob to be Bayesian, then

$$\mathbb{E}[b] = \Pr[s = 1] \cdot b_{s=1} + \Pr[s = 0] \cdot b_{s=0}, \tag{3.1}$$

where

$$b_{s=1} = \frac{b_0\theta}{b_0\theta + (1 - b_0)(1 - \theta)} \quad , \quad b_{s=0} = \frac{b_0(1 - \theta)}{b_0(1 - \theta) + (1 - b_0)\theta} \, ,$$
$$\Pr[s = 1] = a_0\theta + (1 - a_o)(1 - \theta) \text{ , and } \Pr[s = 1] = 1 - \Pr[s = 0].$$

In words, Anne expects Bob's prior b_0 to influence how Bob updates his beliefs for a given signal, while her own prior a_0 to determine the signal frequencies. It is easy to see that when Anne and Bob share the same prior, we recover the fundamental property of Bayesian updating: beliefs are *Martingale*, i.e., information cannot systematically bias beliefs in any direction. This means that Anne's belief about Bob's expected posterior, which is the same as her own posterior belief would be, should be equal to her and his prior.

When Anne and Bob have different priors the situation changes and *any* information is predicted to move Bob's expected posterior closer to Anne's prior. The recent paper by Navin Kartik et al., 2021 shows that Anne expects a Blackwell more informative signal to bring Bob's expected posterior closer to her own prior. To translate this to our setting, say, Bob has access to two information structures that only differ in signal accuracy, i.e., $1 > \theta_1 > \theta_2 > \frac{1}{2}$. Then, Anne expects that both signal structures will move Bob's average posterior closer to her prior compared to what Bob's original prior was. Moreover, she anticipates that the structure with more precise signals, i.e., θ_1, will result in a larger shift and a smaller final disagreement between Anne's prior and Bob's expected posterior.

This result is known as the *Information Validates Prior (IVP)* property. As we argued in the introduction, its significance is broad, spanning many strategic settings studied in Economics and Political Science. It is precisely this result that we set out to investigate empirically in our paper.

3 Experimental Design

Given our interest in how participants think others update their beliefs when they may have potentially different priors we chose to work with genuine, home-grown

beliefs participants have about various facts. In the next section, we discuss the advantages and disadvantages of using this method compared to induced beliefs.

Specifically, we used twelve factual statements in the experiment. Each statement is either true or false. Participants know that the experimenter knows whether the statement is true or false, but naturally may hold different beliefs about the probability that a statement is true. Here are two examples of such statements[9]:

- In 2023, the United States spent more than 10% of the federal budget on foreign aid.
- Rhino horn is made up of keratin—the same protein which forms the basis of our hair and nails.

Treatments. The experiment consists of three main treatments. Treatment T0 is the benchmark treatment, in which we document how people update their own beliefs in response to new information. The purpose of treatment T1 is to study how Anne thinks Bob updates his beliefs when she knows Bob's signal, i.e., Anne's beliefs about Bob's conditional posteriors. Finally, the purpose of treatment T2 is to study Anne's beliefs about Bob's expected posterior, i.e., the situation in which Anne does not observe Bob's signal.

Structure of the experiment. Each treatment consists of three parts. Participants receive the instructions for the next part after they complete the previous one. Instructions before each part include a comprehension quiz to check participants' understanding and focus their attention on the main features of the experiment. The instructions and the screenshots are presented in the Online Appendix.

Part 1 consists of six rounds and is the same in all treatments. In each round, we present participants with one of the statements and elicit their priors about the chance that the statement is true. We then provide participants with a partially informative signal about the correctness of the statement and elicit their posterior about the chance that the statement is true. Signals are generated from two signal structures: a more precise one with $\theta_1 = 0.90$ and a less precise one with $\theta_2 = 0.65$. One of these two structures is randomly selected in each round, and a participant

[9]Figures C.1 and C.2 in the Appendix present all statements used in the experiment and the visualization used alongside the statements.

knows signal accuracy when she makes her choices. We will use these two signal structures to investigate the IVP property which requires comparing the more- and the less-precise information structures. Participants receive no feedback at the end of each round in Part 1. After completing a round, they move on to the next one and are shown the next statement.

Part 2 consists of six rounds as well and is different in each treatment. In T0, Part 2 is the same as Part 1. That is, participants go through another set of 6 statements, report their priors, observe signals, and report their posteriors. A key reason for collecting extensive data on participants' own belief updating is to calculate participants' payments in treatments T1 and T2. This requires observing posteriors for each statement across different signal realizations within various signal structures, which is what we do in T0. We conducted T0 a few days prior to the other treatments to ensure this data would be available for payment calculations.

Part 2 in the remaining two treatments is slightly different. In each round, participants start by observing a statement and reporting their prior. Then, they are matched with past participants from T0 and observe the past participants' prior for the same statement and signal accuracy. Participants in T1 also observe the signal realization received by the past participants, while in T2, no such information is provided. In both treatments, after observing the information, participants are asked to guess the posterior reported by these past participants.[10]

The last part, Part 3, consists of just one round and was administered only to participants who reported a corner prior for one of the statements. If such an event happened, then one of the questions for which a corner prior was reported was chosen and a participant was offered a choice between a very risky bet and a safe payment of \$10. The risky bet pays \$11 in case the reported prior is correct and \$0 if it is wrong. The goal of this final (surprise) round was to gauge how much faith people have in their corner beliefs when they report them. Risking losing \$10 makes sense only if one has little doubt in the reported belief.

At the end of the experiment, participants answered a few unincentivized questions about the difficulty of the experiment. In addition, following McGranaghan et al., 2024, every three rounds, we presented participants with an unincentivized visual brain break to reduce fatigue (see an example in the Online Appendix).

[10] We refer the reader to the Online Appendix for the screenshots detailing the language used to explain these tasks to participants.

Order of statements. Since all rounds are the same in Parts 1 and 2 in T0, the order of statements was randomized across participants in this treatment. For T1 and T2, we split the statements into two batches (batch A consists of statements 1 to 6 and batch B consists of statements 7 to 12). We, then, conducted two versions of each treatment: T1A and T2A used batch A in Part 1 and batch B in Part 2, and T1B and T2B used batch B in Part 1 and batch A in Part 2. Within each part, the order of statements was randomized across participants.[11]

Parameters. Testing the IVP and the Martingale properties requires a variation in Anne and Bob's priors, capturing both similar and distinct priors between the two and spanning a wide range of possible priors. To do so, we match T1 and T2 participants with T0 participants with six pre-selected priors, $b_0 \in \{0.10, 0.20, 0.60, 0, 70, 0.90, 1.00\}$.[12] For each prior b_0, we selected two statements that had a sufficient number of participants reporting such a prior in T0 and providing us with posterior beliefs of past participants (participants in T0) for each signal realization and each signal accuracy. As described above, such data is necessary for computing the payments of participants in T1 and T2.[13]

Subject pool. The experiments were conducted on the Prolific platform in January 2024 with roughly 200 participants in each treatment, for a total of 603 participants. We recruited participants between the ages of 21 and 65, who live in the United States, specify English as their first language, and have a high (90+) approval rating on Prolific. For each treatment, an equal number of men and women were recruited.

Participants' payments. All participants received a fixed payment upon completion: \$3 in the T0 and \$4 in the T1 and T2 treatments.[14] In addition, each participant had a 20% chance to be selected into a bonus group. For the selected participants, the computer randomly chose one of the questions from one randomly selected round for payment. The answer submitted in the chosen question determined whether

[11] This design mitigates the concern that some of the patterns we find in the data are driven by specific statements people saw in one part of the experiment.

[12] Figures C.1 and C.2 in Appendix present the distribution of priors elicited from participants in T0 and indicate which prior was used as past participants' priors in T1 and T2.

[13] An alternative design would be to match participants from T1 and T2 randomly with past participants from T0. The drawback of this design is that an even larger amount of data is required for T0 to ensure that all signal realizations occur for both signal structures and all priors of past participants, some of which are naturally quite rare.

[14] These completion fees are standard, given the average time it takes to complete each treatment.

the selected participant received an additional bonus of $10. We used the standard BDM method to incentivize subjects to truthfully state their beliefs.[15] In addition, in each treatment, we randomly selected eight participants to receive an additional bonus based on their decisions in Part 3 (the corner beliefs). Treatment T0 lasted about 16 minutes and participants earned, on average, $4. Treatments T1 and T2 lasted about 20 minutes and participants earned, on average, $5.

Implementation. The experiment was approved by Caltech (IR21-1179) and pre-registered on aspredicted.org (#158497).[16] The experimental software was programmed in Qualtrics. Instructions and screenshots of the interface are presented in the Online Appendix. Table 3.1 summarizes the details of all three treatments.

Table 3.1: Design

Treatment		Part 1 own beliefs 6 rounds	Part 2 others' beliefs 6 rounds	Part 3 corner beliefs at most 1 round	Nb participants
T0	elicit observe report	own prior signal acc., signal own posterior	own prior signal acc., signal own posterior	risky bet	201
T1	elicit observe report	own prior signal acc., signal own posterior	own prior other's prior, signal acc., signal other's conditional posterior	risky bet	198
T2	elicit observe report	own prior signal acc., signal own posterior	own prior others' prior, signal acc. others' expected posterior	risky bet	202

3.1 Discussion of Experimental Design

In this section, we discuss the rationale behind our key design choice of using genuine, homegrown beliefs instead of inducing beliefs in a neutral context.

[15] The BDM payment is theoretically an incentive-compatible method for eliciting truthful responses regardless of participants' risk attitudes (Becker et al., 1964). In addition, following Danz, Vesterlund, and Wilson, 2021, we told participants that they had no incentive to report beliefs falsely if they wanted to maximize the expected payoff in the experiment. This technique became standard in the literature as it helps participants to understand payment method and, as a result, helps the experimenter to elicit participants' true beliefs.

[16] We conducted two small pilots (pre-registration #110598 and #124788) with different framings of the main belief-updating task to test the software and verify standard behaviors documented in the literature. During this pilot, we identified software errors and realized that our modified framing was unclear to participants. Consequently, we reverted to the standard framing from the literature, focusing on eliciting genuine priors rather than inducing priors. Results from this pilot are available from the authors upon request.

To study the IVP property, one needs an environment in which participants have different beliefs. There are two ways to do that. The first approach involves inducing varying beliefs by providing participants with private signals about the state (Andreoni and Mylovanov, 2012). The second approach prescribes eliciting participants' genuine, naturally formed beliefs about certain factual events (Thaler, 2024).[17]

Both methods have their advantages and disadvantages. The primary advantage of inducing beliefs lies in the ability to control participants' beliefs. This is straightforward when inducing a common belief among all participants. However, it becomes more challenging when inducing heterogeneous beliefs, as this requires participants to update their beliefs based on the private signals they receive. Given the extensive literature documenting deviations from Bayesian updating (Benjamin, 2019), it is unclear whether an experimenter employing this approach can effectively control the induced priors.[18]

Working with genuine beliefs sidesteps this issue, as individuals naturally hold differing beliefs on various topics, including factual statements. Moreover, participants are not likely to be surprised when they learn that others have different views. However, this approach requires collecting a substantial amount of data to capture the variations in Anne's and Bob's priors necessary for evaluating the IVP and Martingale properties.

Our approach of eliciting genuine beliefs as opposed to induced beliefs offers three additional advantages. First, it provides the enhanced external validity of the results, as they directly speak to how people adjust their natural beliefs in response to new information. Second, this approach enables us to investigate whether genuine beliefs about neutral topics—such as general knowledge statements—respond differently to new information compared to politically charged statements. Our study offers

[17]The focus on factual events as opposed to future events that have not happened yet is dictated by the need to incentivize people to report their beliefs truthfully, which requires the experimenter to know the state—in our case, whether the statement is correct or false.

[18]The additional subtle issue with inducing heterogeneous beliefs is what Anne can infer from Bob's prior about Bob's ability to use new information. To illustrate, consider a standard environment with two urns containing balls of different colors. Both Anne and Bob know the compositions of the urns and the chance that each urn is selected; the selected urn represents the state. Each observes a private draw from the urn and forms a belief about the state. These formed beliefs could potentially serve as Anne's and Bob's priors for the investigation of the IVP and Martingale properties. Say, Bob's posterior belief is communicated to Anne. If this belief is unreasonable given the composition of the urns, then Anne will make inferences about Bob's ability to update already at the inducing-the-priors stage of the experiment, which would confound Anne's beliefs about how Bob updates his beliefs given new information.

a preliminary exploration of these differences, and we hope future research will expand it and provide more comprehensive evidence. Third, it allows observing genuine corner priors—instances where participants report extreme confidence in the statement being either true or false. These cases are particularly interesting because they allow us to examine whether corner beliefs are degenerate, as theory suggests, or if they can respond to new information. This type of analysis would not be possible with induced beliefs.

4 Results

We start with presenting reduced-form evidence on Martingale and IVP properties (Sections 4.1 and 4.2). In this analysis, we use the data from T2 treatment, in which Anne predicts Bob's expected posterior. In Section 5 we explore what drives these aggregate results by studying how Anne updates her own beliefs, how Anne thinks Bob updates his beliefs when she knows his signal realization, and what this means for Anne's beliefs about signal distribution.

Approach to Data Analysis. We define Anne and Bob as having the *same priors* if their priors differ by no more than 5 percentage points, and *different priors* if the difference exceeds 5 percentage points. We further categorize the extent of their differences using the following distinctions. We call Anne's and Bob's priors *very polarized* if they differ by more than 40 percentage points, *polarized* if the difference falls between 20 and 40 percentage points, and *somewhat polarized* if the difference is between 5 and 20 percentage points. Statistical tests are performed using regressions, in which we cluster standard errors by individuals to account for the inter-dependency of observations that come from the same participant.[19]

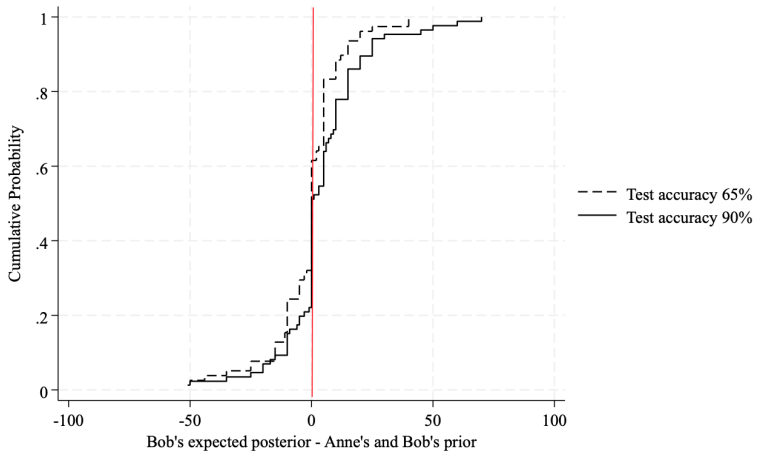
4.1 Martingale property

Does Anne expect information to systematically alter Bob's posterior when they share prior? Figure 3.1 presents the CDFs of the difference between Anne's beliefs

[19]To be precise, we regress the variable of interest (for instance, the difference between Anne's prediction about Bob's expected posterior and Anne's prior) on a constant and an indicator for one of the treatments (for instance, the test accuracy), while clustering standard errors at the individual level. We say that the two treatments are significantly different when the estimated treatment indicator is different from zero at the standard 5% level and report the p-value associated with the estimated indicator.

about Bob's expected posterior and his original prior, which also happens to be her prior.

Figure 3.1: Changes in Bob's beliefs when Anne and Bob share the same prior



Notes: The difference between Anne's prediction about Bob's expected posterior and Anne's own prior is reported. We focus on cases in which Anne and Bob have similar priors. The data is from Part 2 of treatment T2.

The average difference is not significantly different from zero when the test accuracy is 65% ($p = 0.507$) and is significantly different from zero but very small when the test accuracy is 90% (equals 4 percentage points, $p = 0.051$). Moreover, as seen from Figure 3.1, the two CDFs are close to be symmetrically distributed around zero, and there is no statistical difference across the two signal structures ($p = 0.135$). This analysis provides strong support for the Martingale property.

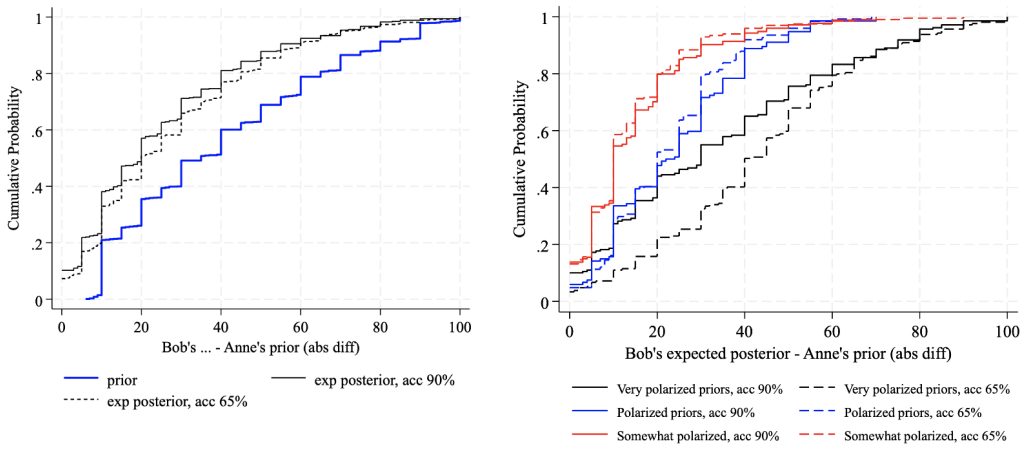
Observation 1: Anne believes that Bob's beliefs satisfy the Martingale property, i.e., from an ex-ante perspective, information cannot alter Bob's beliefs.

4.2 IVP property

What does Anne think about Bob's expected posterior when they have different priors? The IVP property has two empirical footprints. First, Anne expects any information to be effective at bringing Bob's posterior closer to her prior relative to the original disagreement in their priors. Second, the more precise information is expected to decrease disagreements between Anne and Bob by producing larger shifts in Bob's expected posterior relative to the less precise information. Table 3.2 and Figure 3.2 depict the basic statistics and cumulative distribution functions

(CDFs) of the absolute differences between Bob's expected posteriors and Anne's priors for the two signal structures, as well as the original difference in opinions (priors) between them.

Figure 3.2: Changes in Bob's beliefs when Anne and Bob have different priors



Notes: The left panel depicts the CDFs of the absolute differences between Bob's and Anne's priors, as well as the absolute differences between Bob's expected posteriors and Anne's priors. The right panel displays the differences between Bob's expected posteriors and Anne's priors, broken down by each level of prior disagreement. The analysis in both panels focuses on cases where Anne and Bob have different priors.

Consistent with the IVP prediction, any information structure moves Bob's posteriors closer to Anne's priors. The shift is large in magnitude and significant at 1% level regardless of how different Anne's and Bob's original opinions were.[20] Moreover, more precise signals shift Bob's beliefs closer to those of Anne. In fact, the CDF curve for test accuracy 65% first-order stochastically dominates the one for test accuracy 90%. However, the difference between the two CDFs is rather small and only marginally significant ($p = 0.079$).

The right panel in Figure 3.2 and the data in Table 3.2 show that the difference between the two information structures primarily comes from cases in which Anne's and Bob's original priors are very polarized (at least 40 pp apart). Put differently, when Anne and Bob have very different initial opinions, Anne expects Bob's posterior to move closer to her prior when he learns from a more accurate source. The effect in this case is large in magnitude and highly significant ($p = 0.001$): the

[20] The CDFs of the differences between Bob's and Anne's priors for two signal structures overlap. This is by design, as Bob's signal precision was randomly assigned in the experiment. For brevity, this graph is omitted but is available from the authors upon request.

Table 3.2: Differences in Anne's and Bob's beliefs before and after Bob consumes new evidence, in absolute terms.

Bob's prior is different from Anne's prior (at least 5 pp difference)

			Anne's prior					
	all		extreme		intermediate		close to uniform	
	mean (se)	med	mean (se)	med	mean (se)	med	mean (se)	med
before info	39.8 (1.0)	35	44.4 (1.5)	40	35.6 (1.1)	30	30.6 (1.2)	35
info acc 90%	24.8 (1.1)	20	24.5 (1.6)	15	25.0 (1.5)	20	24.6 (1.6)	30
info acc 65%	27.3 (1.0)	20	31.1 (1.5)	25	24.0 (1.5)	20	20.7 (1.9)	20
p-values								
before vs 90%	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p = 0.018$	
before vs 65%	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
90% vs 65%	$p = 0.079$		$p = 0.002$		$p = 0.440$		$p = 0.098$	

Anne's and Bob's priors are very polarized (more than 40 pp difference)

			Anne's prior					
	all		extreme		intermediate		close to uniform	
	mean (se)	med	mean (se)	med	mean (se)	med	mean (se)	med
before info	67.9 (0.8)	65	75.0 (0.9)	70	55.9 (0.8)	55	49.6 (0.7)	50
info acc 90%	34.0 (2.0)	30	36.1 (2.7)	30	28.5 (2.8)	25	n/a	n/a
info acc 65%	42.4 (1.8)	40	47.2 (2.2)	49	37.0 (2.4)	40	29.5 (4.6)	33
p-value								
before vs 90%	$p < 0.001$		$p < 0.001$		$p < 0.001$		n/a	
before vs 65%	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
90% vs 65%	$p = 0.001$		$p = 0.001$		$p = 0.024$		n/a	

Notes: Each table reports the difference between Anne's and Bob's prior beliefs in the first row and the differences between Anne's beliefs about Bob's expected posterior and her own prior in the second and third rows. The second and third rows differ by signal accuracy. We focus exclusively on cases where Anne and Bob have different priors. Entries marked n/a indicate instances with fewer than 10 observations. Anne's prior is categorized into three groups: extreme priors (below 20 or above 80), close-to-uniform priors (between 40 and 60), and intermediate priors (the remaining category, i.e., priors between 20 and 40 or between 60 and 80).

median shift is 10 pp and it is almost 20 points when Anne has extreme priors. At the same time, contrary to the IVP property, when Anne's and Bob's priors differ by less than 40 pp, we observe no difference between the two information structures in general (the two blue and the two red lines in the right panel of Figure 3.2 are very similar).[21] Overall, when Anne's and Bob's opinions are not too different, Anne

[21] Table C.1 in Appendix, presents similar statistics as the bottom part of Table 3.2 for the case in which Anne's and Bob's priors differ by at most 40 pp. The data shows that when Anne's beliefs are extreme, the more precise information structure shifts Bob's expected posterior closer to Anne's prior, consistent with the IVP property. However, the opposite is true when Anne's priors are intermediate or close to uniform; in these cases, Anne thinks that the less precise signals are more effective at

believes that Bob's posteriors will shift similarly regardless of whether he learns from a more precise or a less precise source.

Anne's beliefs about changes in average Bob's beliefs for politically charged statements are similar to those for non-politically charged statements but with an even greater disregard for the quality of information compared to neutral statements. Figure C.3 in Appendix replicates Figure 3.2 for two politically charged statements used in our experiment.[22] It illustrates that Anne believes Bob's average posterior beliefs will still move closer to her own prior after receiving new information. Yet, unlike neutral statements, these shifts are identical regardless of the quality of the information Bob receives.

How do Anne's estimates about shifts in Bob's posteriors compare to those predicted by Bayesian theory? Figure 3.3 combines the data from both information structures and depicts the difference between predicted and observed posteriors depending on whether Anne's prior is above or below that of Bob's and how different the two priors are (black solid and dashed lines).[23] As a reference, we also plot the case in which Anne and Bob share the same prior (red lines).

As we've discussed above, when Anne shares the same prior as Bob, she expects Bob's posterior to be on average the same as Bob's or her own prior. In Figure 3.3 this is depicted by red lines being very close to zero and actually being exactly equal to zero for a large portion of the data. However, when Anne and Bob do not share the same priors, Anne tends to underestimate how Bob's posteriors move relative to those predicted by Bayesian theory. Indeed, when Anne's prior is above that of Bob, the majority of Anne's estimates are below predicted ones, while the opposite happens when Anne's prior is below that of Bob. Figure C.6 in Appendix replicates Figure 3.3 for Anne's extreme beliefs, for which we find the strongest qualitative support of the IVP, and document similar patterns. Overall, Anne expects that Bob's reaction to information is much more inert than what the Bayesian theory predicts.

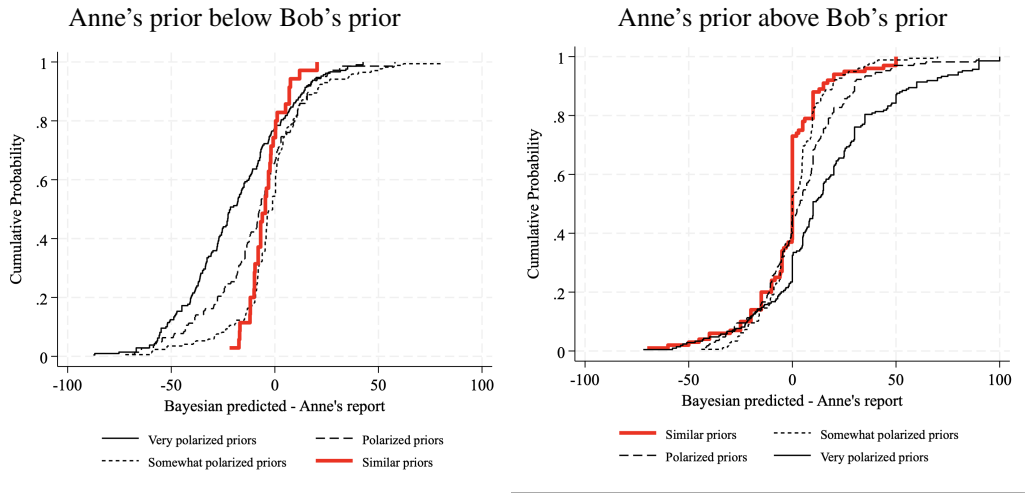
Observation 2: We find partial support for the IVP property. Consistent with the IVP, Anne thinks that any information brings Bob's average opinion closer to her own,

reducing the polarization of opinions between Anne and Bob.

[22] We have two politically charged statements: statement 6 about the estimates of GDP growth under Democratic vs Republican presidents and statement 3 about the United States foreign aid spendings.

[23] Figure C.5 in Appendix presents Figure 3.3 separately for each information structure and shows very similar patterns.

Figure 3.3: Anne's estimates of Bob's expected posteriors vs Bayesian predictions



Notes: We plot the CDFs of the differences between Bob's Bayesian-predicted posterior expectations and Anne's estimates of these values. The plots are separated into cases where Anne's prior is lower than Bob's (left panel) and cases where Anne's prior is higher than Bob's (right panel). The data is sourced from Part 2 of treatment T2.

and more precise information is (marginally) more effective at this job. However, the significant difference in the effectiveness of more precise information structures is observed only when Anne has very different beliefs from Bob, when Anne holds relatively extreme prior beliefs herself, and when the statements are neutral. Otherwise, Anne expects Bob's beliefs to shift similarly regardless of information quality. In other words, Anne thinks that Bob's average posteriors are not responsive to information quality, contrary to what the Bayesian theory predicts.

5 Unpacking Aggregate Results

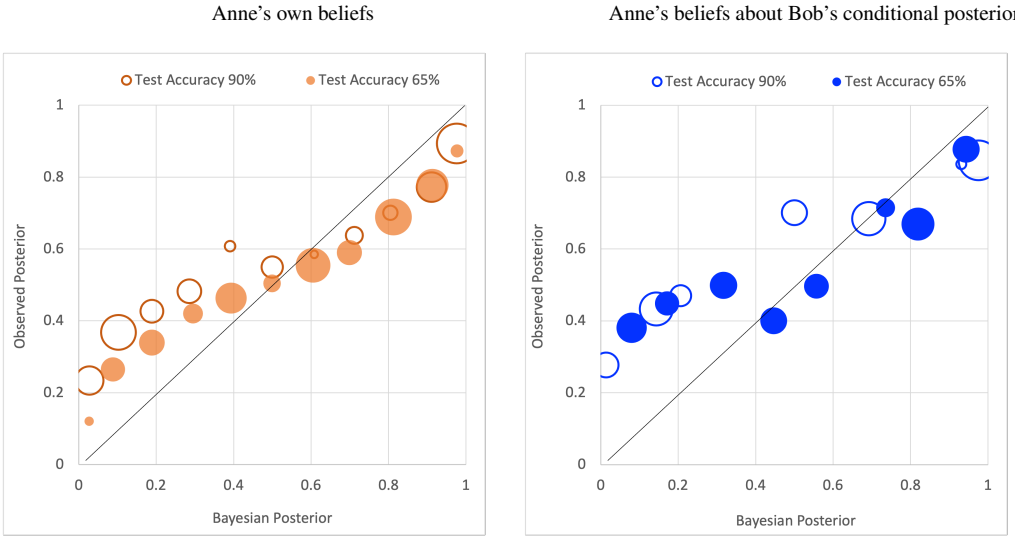
In this section, we study what drives aggregate results presented in Section 4. We start by documenting how Anne revises her own beliefs in response to new information using the data collected in treatment T0 and in Part 1 of treatments T1 and T2 (Section 5.1). We then investigate how Anne thinks Bob forms his conditional posteriors when new information arrives using the data collected in Part 2 of treatment T1 (Section 5.2). After that, in Section 5.3, we study how Anne predicts the signal frequencies based on what we learned about Anne's updating process in Section 5.1. Section 5.4 closes the loop by bringing all these elements together and explaining why Anne's beliefs about Bob's expected posteriors are much more rigid than Bayesian theory predicts.

The analysis is structured as follows. We first present model-free raw data patterns. Then, we analyze the data through the lens of the Grether, 1980 model, which is one of the most popular behavioral models used in the literature to account for deviations in belief-updating tasks (Benjamin, 2019). In our case, it turns out that the Grether model outperforms alternative behavioral models proposed in the literature. We discuss these alternatives and run a horse race between the models in Appendix C.1.

5.1 How Anne Updates Her Beliefs

The left panel in Figure 3.4 depicts Anne's reported posteriors as a function of Bayesian posteriors. For both signal accuracies, we observe a familiar inverse S-shape (Benjamin, 2019; Enke and Graeber, 2023). People tend to overestimate the probabilities of unlikely events and underestimate the probabilities of very likely events.

Figure 3.4: Observed versus Bayesian posteriors



Notes: The left panel uses data from treatment T0 as well as Part 1 from treatments T1 and T2, while the right panel uses the data from Part 2 of treatment T1. In both panels, we exclude degenerate corner priors.

Grether, 1980 model proposes a parsimonious way to modify Bayes's rule which allows accommodation of over- and under-inferences from either or both the prior and the signals. This model is parameterized by parameters (c, d) which captures the degree to which updating deviates from the Bayesian one. Specifically, Anne's

posterior given signal $s = 1$ can be written as

$$a_{s=1} = \frac{a_0^c \theta^d}{a_0^c \theta^d + (1 - a_0)^c (1 - \theta)^d},$$

The model collapses to the Bayes's rule when $c = d = 1$. Otherwise, the parameter c controls the weight on the prior, and the parameter d controls the weight on the new information. Both parameters matter in determining how sensitive Anne's posterior is to her initial beliefs and newly received signals.

Table 3.3: Anne's Own Posteriors and Anne's Beliefs about Bob's Conditional Posteriors, estimates of Grether model

	Dependent Variable = ln [Posterior odds]					
	Anne's own beliefs		Anne's beliefs about Bob's conditional beliefs			All together
	reg (1)	reg (2)	reg (3)	reg (4)	reg (5)	reg(6)
ln [Prior odds]	0.55** (0.02)	0.55** (0.02)	0.31** (0.04)	0.27*** (0.04)	0.31** (0.04)	0.29** (0.04)
ln [Likelihood ratio]	0.46** (0.02)	0.47** (0.02)	0.43** (0.04)	0.43*** (0.04)	0.47** (0.04)	0.42** (0.04)
ln [Prior odds]						
x Political		-0.02 (0.05)			0.70** (0.32)	
ln [Likelihood ratio]						
x Political		-0.09** (0.04)			-0.17** (0.09)	
ln [Prior odds] *x* Anne						0.26** (0.04)
ln [Likelihood ratio]						
x Anne						0.04 (0.04)
ln [Prior odds]						
x Same Priors				0.38*** (0.08)		
ln [Likelihood ratio]						
x Same Priors				0.04 (0.08)		
Nb obs	$n = 3534$	$n = 3534$	$n = 865$	$n = 865$	$n = 865$	$n = 4261$
Nb participants	$i = 581$	$i = 581$	$i = 195$	$i = 195$	$i = 195$	$i = 582$
R-squared	0.4433	0.4443	0.2956	0.3116	0.3068	0.4193
Data	both parts in T0 Part 1 in T1 and T2		Part 2 in T1			both parts in T0 both parts in T1 Part 1 in T2

Notes: We express Grether's formula in the log form, i.e., $\ln \frac{a_{s=1}}{1-a_{s=1}} = c \cdot \ln \frac{a_0}{1-a_0} + d \cdot \ln \frac{\theta}{1-\theta}$. This implies a linear relationship between the posterior odds, the prior odds, and the likelihood ratio. We estimate this relationship using linear regression with the standard errors clustered at the individual level. We exclude Anne's degenerate priors in reg (1), (2), and (6) and Bob's degenerate priors in reg (3) - (6). Political is an indicator of two politically charged statements (the GPD growth and the foreign aid spending). Anne is an indicator of Anne's own posteriors. Same is an indicator that Anne's and Bob's priors are within 5 p.p. from each other. ** indicates significance at the 5% level.

Regression (1) in Table 3.3 reports estimates of parameters (c, d) for Anne's own beliefs when she receives new information. Our results are consistent with the canonical findings in the literature established for the belief-updating tasks with

so-called balls-and-urns experiments and induced beliefs: both parameters c and d are significantly smaller than the Bayesian benchmark (Benjamin, 2019). This means that people tend to under-infer from both the new information they receive and their own homegrown genuine priors. Regression (2) distinguishes between neutral and politically charged statements and shows that people put similar weight on their priors in both cases but update less in light of new evidence related to political statements.

Anne's Corner Beliefs. Corner beliefs are degenerate, and, by definition, unchangeable. If you are absolutely sure that a statement is either true or false, then no new evidence should alter your conviction, and your opinion of the statement should remain unchanged. Our experiment provides one of the first empirical evidence evaluating this prediction.[24]

How often do people report corner beliefs? That depends on the statement. The fraction of corner beliefs ranges from 11% to 49% per statement, with an average of 22%. Some participants are more likely to report the corner beliefs than others. However, as Figure C.4 in Appendix shows, participants rarely report corner beliefs for more than 4 statements out of 12 in total.[25]

Do people take corner beliefs seriously? The data in Part 3 of the experiment provides some insights into this question. Recall that in this part, we offer participants a choice between a safe payment of \$10 and a risky bet which pays \$11 if one's reported corner belief is correct and nothing otherwise. About three-quarters of participants who reported a corner belief chose the risky bet in the last part of the experiment. We take this evidence as supportive of the fact that people do originally believe in their corner priors. Taking such a risky bet makes sense only if one has little doubt about correctly assessing the truthfulness of the statement.[26]

[24]Note that, by design, both signals are conceivable even when one holds a corner prior. This is true because the signals are only partially informative: conditional on the state, there is a positive chance of receiving either a signal that coincides with the state or contradicts it. Thus, a participant cannot learn from a signal that their prior is wrong.

[25]Recall, that Figures C.1 and C.2 present the histograms of prior beliefs for each statement observed in T0. Participants' priors in the other two treatments T1 and T2 are very similar to those in T0 and are omitted for brevity.

[26]Focusing on participants whose last surprise round involved the statement where they reported a corner belief and received a signal about that statement—i.e., the 'own beliefs' portion of the experiment—we find that they are more likely to choose the risky bet when the signal confirms their prior belief than when it contradicts it. Specifically, participants with confirming signals chose the risky bet over 80% of the time, compared to approximately 65% for those with contradicting signals.

Do people update corner beliefs? Figure 3.5 depicts the CDFs of Anne's posteriors after receiving either a confirming original prior signal (left panel) or a contradicting original prior signal (right panel). We pool the data from both corners and redefine all corner beliefs to be one. The confirming signal is, then, a more likely signal conditional on the state being one, while the contradicting signal is the less likely signal.[27] The red thick lines depict updating for corner beliefs, while the black solid and dashed lines provide a benchmark of how Anne updates her beliefs when her prior is close to the corner but still interior.

When Anne receives a confirming signal, she rarely updates (left panel of Figure 3.5). The median posterior, in this case, is 100, the average is 92, and it is significantly different from Anne's posterior when she has a slightly lower prior between 90% and 99% and similarly receives a confirming signal ($p < 0.001$). However, when Anne receives a contradicting signal (right panel of Figure 3.5), she updates her beliefs substantially. The median belief in this case is 80, the average is 63, and it is not significantly different from the posterior beliefs of Anne whose prior is between 90% and 99% and who similarly receives a contradicting signal ($p = 0.152$). This evidence suggests that corner beliefs are not really corners: people are willing to change their minds in light of new evidence that goes against their prior beliefs, even when they were initially certain in their opinion.

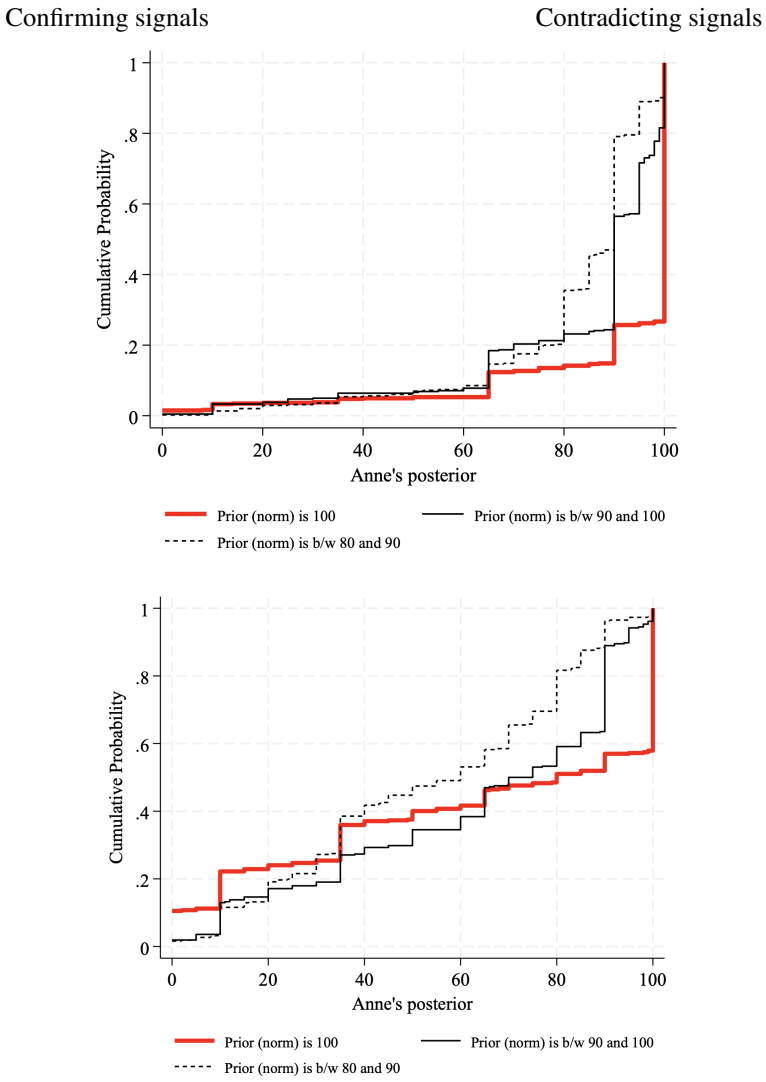
Observation 3: When updating her own beliefs, Anne underinfers both from her prior and from new information. The underinference from new information is stronger for politically charged statements. Corner beliefs are not really degenerate, they are malleable to some degree and can be updated in light of contradictory evidence.

5.2 How Anne Thinks Bob Updates His Beliefs

The right panel in Figure 3.4 depicts Anne's guesses about Bob's conditional posteriors as a function of Bayesian posteriors. For both signal accuracies, we observe a familiar inverse S-shape similar in form to Anne's own posteriors (left panel). The shape similarity between the left and the right panels in Figure 3.4 is consistent with Anne projecting her way of updating onto how she thinks others update, albeit larger deviations from the Bayesian predictions for Anne's own posteriors compared

[27] Say, a participant believes that the statement is correct with probability 100% and receives a signal which is 90% accurate. The positive signal is a confirming signal, as it confirms the original belief of a participant. The negative signal, however, is a contradicting one, as it goes against one's prior.

Figure 3.5: How Anne updates her own corner beliefs



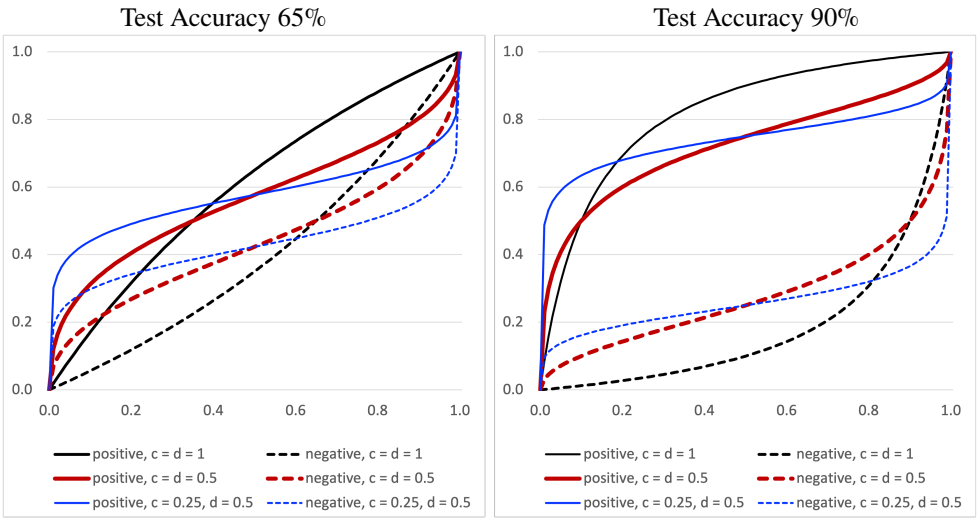
Notes: We plot the CDFs of Anne's normalized posteriors of Anne from Part 1 in all treatments. The normalized prior equals 100—elicited prior for priors below 50 and equals itself for the priors above 50. The confirming signal is a more likely signal and the contradicting signal is the less likely signal conditional on the statement being true (100% correct).

to Bob's posteriors.[28] Regression (3) presented in Table 3.3 confirms what we see in Figure 3.4: Anne thinks that Bob, like her, underinfers both from new evidence and from his prior, i.e., $c^{\text{Bob}} < 1$ and $d^{\text{Bob}} < 1$. Regression (6) shows that Bob's

[28] Loewenstein et al., 2002 show that people project their current tastes on their future selves. Danz, Madarasz, and Wang, 2024 show evidence that people project their own biases on others. We document that people project the way they update on how others do so when encountering new evidence.

underinference from the prior is stronger than her own, i.e., $c^{\text{Bob}} < c^{\text{Anne}}$.

Figure 3.6: Bob's conditional posteriors after receiving a signal as a function of his prior



Notes: Each panel depicts Bayesian posteriors and Grether's posteriors for two signal realizations conditional on the parameters (c, d). The left picture is for weak signals (accuracy 65%), while the right is for strong ones (accuracy 90%). Both figures show how underinference flattens posteriors. First, it reduces differences between signal realizations (difference between black lines vs. between red lines). Second, lower c makes posteriors less sensitive to priors (red vs. blue lines, holding d fixed).

The two underinferences (from the prior and from the signals) are important as they compress Bob's conditional posteriors towards a 50/50 belief and result in Anne believing that Bob's posteriors are not as responsive to Bob's priors as Bayesian theory predicts or as Anne's own beliefs respond. Figure 3.6 illustrates this point and plots Bayesian posteriors as well as posteriors predicted by the Grether model for $(c, d) = (0.5, 0.5)$ and $(c, d) = (0.25, 0.5)$. The latter parameters correspond roughly to those estimated in regression 6 of Table 3.3, which are based on Anne's observed beliefs about Bob's conditional posteriors. Both figures illustrate the compression and flattening effects previously discussed. First, the general tendency of individuals to underinfer from new information and priors flattens conditional posteriors relative to Bayesian predictions. This substantially reduces the predicted difference in posteriors for two signal realizations (compare the black lines to the red lines). Second, stronger underinference from one's prior—represented by a lower parameter c leads—to conditional posteriors that are even less responsive to the prior (compare the red and the blue lines, holding d fixed). Both effects are crucial

for understanding why the quality of information has a limited impact on Anne's beliefs about Bob's expected posteriors.

Four more patterns regarding Bob's conditional posteriors are worth noting. First, regression (5) in Table 3.3 distinguishes between the types of statements Bob encounters. Interestingly, Anne thinks that Bob puts a significantly higher weight on his prior and a significantly lower weight on the new evidence for politically charged statements relative to the neutral ones. In other words, Anne believes that relative to the neutral statements, Bob's posteriors regarding political statements will be close to his priors and new information will not have much effect on these priors.

Second, another manifestation of projection bias is evident in Anne's susceptibility to base-rate neglect and her predictions about Bob's likelihood of exhibiting the same bias. The base-rate neglect in its pure form suggests that a decision-maker completely ignores their prior and, for instance, after receiving a positive signal with 90% accuracy updates their posterior to 90%.[29] Using individual-level data, we find a strong and significant correlation between the number of questions where Anne exhibits perfect base-rate neglect and the number of questions where she believes Bob will do the same: corr = 0.59 ($p < 0.01$).

Third, Bayesian theory posits that Anne's beliefs about Bob's conditional posteriors are independent of her prior. Regression (4) shows that this prediction does not match our data as Anne thinks that Bob will put more faith in his prior when he shares the same prior as she does.

Fourth, Anne also projects her own way of updating corner beliefs on Bob. Anne thinks that Bob's corner beliefs are not set in stone, especially when he receives a signal that contradicts his initial prior. Figure C.7 in the Appendix shows trends similar to those in Figure 3.5, illustrating the similarity between how Anne updates her corner beliefs and what Anne thinks about Bob's updating his corner beliefs. There is minimal updating when the signal aligns with the initial prior, but significant adjustment when the signal contradicts it.

Observation 4: Anne projects the way she updates her beliefs on the way she thinks others do. Anne thinks that Bob underinfers both from his prior and from new evidence and that Bob's corner beliefs may shift, just like hers. Compared to her

[29] Base-rate neglect is one of the most prevalent biases in decision-making, garnering considerable attention in the literature due to its persistence and widespread occurrence (Benjamin, 2019; Esponda et al., 2023; Gneezy et al., 2023).

own updating process, Anne thinks Bob underinfers from his prior to a larger degree than she does herself. For political statements, Anne believes that new information is quite ineffective at moving Bob's beliefs, and his posteriors will not differ much from his priors. Overall, Anne expects Bob's posteriors to be less responsive to both his prior and information quality compared to what the Bayesian theory predicts.

5.3 What Anne Thinks about Signal Distribution

The last element in the equation determining Bob's expected posteriors is signal distribution. In the Bayesian world (Section 2), the likelihood of receiving a positive signal depends linearly on Anne's prior belief and the signal accuracy:

$$\Pr[s = 1] = a_0\theta + (1 - a_0)(1 - \theta) \tag{3.2}$$

However, the evidence presented so far documents systematic deviations from the Bayesian model. How do these deviations affect Anne's beliefs about signal frequencies? This is what we study in this section.

Note that understanding what Anne thinks about signal frequencies is an inference exercise, since we do not directly observe Anne's beliefs about signal distribution.[30] We therefore proceed as follows. We start by formulating several alternative behavioral models that describe how Anne may form beliefs about signal frequencies. The models we consider are either based on the behavioral patterns documented above or are popular models in the literature relevant to our setting. We examine the general properties of these models and compare their predictions with those of the Bayesian model. In Section 5.4, we estimate these models using data from treatment T2 and evaluate their ability to fit the data.

The first model is that of Grether, 1980. As documented in Section 5.1, Anne tends to underinfer both from her prior a_0 and from signals, the accuracy of which is depicted by parameter θ. Grether's model is summarized by two parameters (c, d) and it does a good job at tracking the deviations of Anne's beliefs from the Bayesian ones. Applying this model to signal frequencies requires some normalization to

[30] One could envision an experiment in which Anne's beliefs about signal frequencies are elicited, in addition to her beliefs about Bob's average posteriors. However, we chose not to pursue this approach out of concern that it might be leading and could alter how people naturally think about others' average posteriors. For instance, consider someone who does not instinctively break down Bob's average posterior into signal frequencies and Bob's conditional posteriors. If asked a question that prompts this decomposition, they might learn to focus on signal frequencies as a crucial element, even though they might not have done so on their own without such a suggestion.

guarantee that both frequencies are bounded between zero and one and sum up to one. Incorporating these restrictions, we arrive at

$$\Pr[s = 1] = \frac{a_0^c \theta^d + (1 - a_0)^c (1 - \theta)^d}{a_0^c \theta^d + (1 - a_0)^c (1 - \theta)^d + a_0^c (1 - \theta)^d + (1 - a_0)^c \theta^d} \tag{3.3}$$

A crucial feature of the Grether and the Bayesian models is that Bob's prior does not play a role in determining signal frequencies; the latter is solely based on Anne's prior and signal accuracy.

The second model we consider differs from the Grether and Bayesian models in that it allows for *social exchange*. Anne, upon observing that Bob holds a different prior from her own, takes this into account and revises her prior to $\tilde{a}$. She then formulates signal frequencies as suggested by the Bayesian model in equation 3.2, using the revised prior $\tilde{a}$ instead of her original prior a_0.

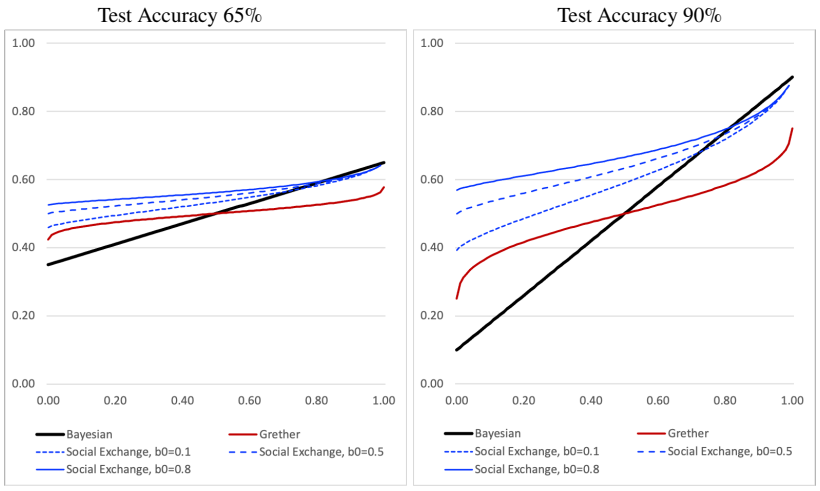
We follow the model of social exchange by Yuksel and Oprea, 2022 to define how Anne revises her prior after observing Bob's prior. This model suggests that Anne takes Bob's prior b_0 at 'face value': Anne consider b_0 to be generated with probability b_0 if the statement is true, and with probability $1 - b_0$ if the statement is false. That is, Anne believes that Bob's prior b_0 is an additional signal about the state of the world, i.e., the truthfulness of the statement, with a likelihood ratio being $\frac{b_0}{1-b_0}$. Thus, we can express the revised prior odds ratio as

$$\log \frac{\tilde{a}}{1 - \tilde{a}} = \alpha \cdot \log \frac{a_0}{1 - a_o} + \gamma \cdot \log \frac{b_0}{1 - b_0}, \tag{3.4}$$

Parameters (α, γ) govern the weight that Anne puts on her prior relative to Bob's prior, and can be estimated from the collected data. If Anne was fully Bayesian, then $\alpha = 1$ and $\gamma = 0$, indicating that Anne is fully confident in her prior and learns nothing from Bob's prior. If, on the contrary, $\gamma > 0$ then Anne adjusts her prior after observing Bob's prior.

While the Grether and the Social Exchange models are obviously different, they share two properties. First, both flatten signal frequencies with respect to Anne's own prior relative to the steepness embedded in the Bayesian benchmark. Second, for a fixed Anne's prior, both reduce the difference in signal frequencies across more and less precise signals. Figure 3.7 demonstrates this point by plotting Anne's estimation of the probability that Bob will receive a positive signal as a function of Anne's prior. The left panel focuses on signals with low precision, $\theta = 0.65$, and

Figure 3.7: Signal Frequencies



Notes: For each behavioral model, we plot the probability that Anne assigns to Bob receiving a positive signal as a function of Anne's own prior. For Grether's model, we use $c = d = 0.5$. For the social exchange model, we use weight $\alpha = 0.75$ on Anne's own prior and weight $\gamma = 0.25$ on Bob's prior. We compute Anne's revised beliefs given these weights and plot three lines: the solid line is for Bob's high prior $b_0 = 0.8$, the dashed line is for Bob's intermediate prior $b_0 = 0.5$, and the dotted line is for Bob's low prior $b_0 = 0.1$.

the right one on signals with high precision, $\theta = 0.9$. In both panels, the black lines depict the Bayesian benchmark, the red lines are for the Grether model, and the blue lines are for the Social Exchange model.

The two effects described above and depicted on Figure 3.7 are important for the following reason. In the Bayesian world, the substantial difference in Bob's expected posteriors for signals of different quality is driven by the significant difference in the *likelihood of receiving positive and negative signals* in two information structures. This is captured by a large difference in the slopes of the two black lines across panels in Figure 3.7. Contrary to that, in Grether model, the difference between the two red lines is significantly smaller and not very responsive to Anne's prior. This means that Anne expects Bob to receive signals with similar likelihoods regardless of whether he is exposed to a high- or a low-accuracy information structure and regardless of her own prior. The same conclusion follows from examining the Social Exchange model (the blue lines).

Observation 5: Compared to the Bayesian benchmark, Anne expects signal frequencies to be less responsive to her own prior and the quality of information Bob consumes. This conclusion holds regardless of the behavioral model Anne uses to

formulate signal frequencies.

5.4 Bringing All Pieces Together

In Section 4.2, we have documented partial support for the IVP property. Consistent with this property, Anne predicts that, in general, information will bring Bob's expected posterior closer to her own prior. However, in contrast with this property, Anne predicts that these posterior moves are similar for information sources with different accuracy. The analysis of Anne's own updating process and her beliefs about Bob's updating process presented in the previous sections helps us understand why this might be the case. We identified two "flattening effects" that jointly reduce the disparity in average posteriors predicted for signals of varying strength. The first flattening effect reflects Bob's diminished sensitivity to conditional posteriors, while the second captures the reduced responsiveness of signal frequencies to both signal accuracy and Anne's prior beliefs. Together, these effects result in Bob's expected posteriors remaining relatively stagnant, showing limited responsiveness to the quality of information he encounters.

We've discussed two behavioral models that can account for the effect of 'average posteriors being less responsive to information quality than predicted by the Bayesian model'. Both models use Grether's framework to calculate Anne's and Bob's posteriors conditional on signal realizations. The two models differ in how signal frequencies are computed; the fully Grether's model uses Grether's logic to compute signal frequencies, while the Social Exchange model allows Anne to revise her prior after observing Bob's prior before formulating signal frequencies.

Table 3.4 estimates both models and compares their fit to the Bayesian benchmark. We perform this exercise twice: once using all observations and modifying corner priors to close but non-corner values (the top part) and once excluding the corner priors (the bottom part). The modification of corner beliefs is warranted given our analysis of how people update their corner priors, which shows that corner priors are not degenerate and can change as new evidence arrives.[31]

To compare the fit we run a simple linear regression of observed posteriors on the predicted ones, clustering standard errors at the individual level, i.e.,

$$\text{Observed Posterior} = \beta_0 + \beta_1 \cdot \text{ Predicted Posterior} + \epsilon. \tag{3.5}$$

[31] Similar modification is done in the analysis of Enke and Graeber, 2023.

The best fit is achieved when $\beta_0 = 0$ and $\beta_1 = 1$.

Table 3.4: Parameter Estimates and Model Fit for Behavioral Models

	Anne's parameters $(c^{\text{Anne}}, d^{\text{Anne}})$	Bob's parameters $(c^{\text{Bob}}, d^{\text{Bob}})$	Revised prior (α, γ)	Model fit β_0	β_1	root MSE
All data						
Bayesian				0.28** (0.01)	0.55** (0.01)	0.2522
Grether	(0.36,0.43)	(0.30,0.47)		0.08** (0.01)	0.94** (0.02)	0.2461
Grether + Social Exchange	(0.39,0.40)	(0.26,0.45)	(1,1)	0.14** (0.01)	0.82** (0.02)	0.2517
Without corners						
Bayesian				0.27** (0.01)	0.57** (0.01)	0.2303
Grether	(0.49,0.41)	(0.33,0.43)		0.07** (0.01)	0.96** (0.02)	0.2288
Grether + Social Exchange	(0.48,0.40)	(0.32,0.46)	(1,0.58)	0.06** (0.01)	0.97** (0.02)	0.2278

Notes: The estimates of each model are presented alongside the model fit (see equation 3.5). We use data from all parts of all treatments. The top part of the table uses all data including the corner priors, where corner priors of 100% and 0% are replaced by 99% and 1%, respectively. The bottom part of the table excludes these corner priors.

Two patterns emerge from Table 3.4. First, the estimated parameters (c, d) for Anne's own updating and Anne's beliefs about Bob's updating are similar to those presented in Sections 5.1 and 5.2, and display the same qualitative patterns. In particular, Anne thinks that Bob's under-inference from new evidence is similar to her own, i.e., $d^{\text{Anne}} = d^{\text{Bob}}$. At the same time, Anne thinks that Bob under-infers from his prior more than she does herself, i.e., $c^{\text{Anne}} > c^{\text{Bob}}$.[32]

Second, both models perform very well at explaining deviations from the Bayesian benchmark. Among the two of them, we favor Grether's model since it uses fewer parameters than the model that combines elements of Grether's model and the Social Exchange and has a similar or better fit.

Magnitudes of Two Flattening Effects. Here, we measure the relative importance of the two flattening effects. We do that through the prism of Grether's model which, as we've argued above, is an elegant and parsimonious way of organizing our data.

Table 3.5 performs a decomposition exercise and turns on/off the two flattening effects one at a time. The first row is the Bayesian benchmark, where both the

[32] We cannot reject the hypothesis that $d^{\text{Anne}} = d^{\text{Bob}}$. We obtain $p = 0.24$ ($p = 0.55$) for Grether model using all data (data without corners). We obtain $p = 0.10$ ($p = 0.11$) for Grether + Social Exchange model using all data (data without corners). At the same time, we reject the hypothesis that $c^{\text{Anne}} = c^{\text{Bob}}$ in all specifications of all models ($p < 0.01$).

Table 3.5: Decomposition of the Combined Flattening Effect

	Conditional Posteriors	Signal frequency	Model Fit β_0	β_1	root MSE
(1)	Bayesian	Bayesian	0.28** (0.01)	0.55** (0.01)	0.2522
(2)	Bayesian	Grether	0.19** (0.01)	0.73** (0.02)	0.2553
(3)	Grether	Bayesian	0.07** (0.01)	0.96** (0.02)	0.2440
(4)	Grether	Grether	0.08** (0.01)	0.94** (0.02)	0.2461

Notes: We use all the data from all treatments and modify corner priors from 100% and 0% to 99% and 1%, respectively. The results are similar when we exclude corner priors (see Table C.2 in Appendix).

signal frequencies and Bob's conditional posteriors are assumed to be Bayesian. This model is a good benchmark but does not fully account for behavioral patterns observed in our experiments. Allowing either signal frequencies or conditional posteriors to follow Grether's model improves the fit significantly, with the latter modification outperforming the former one. The last row is Grether's model, where both elements follow Grether's logic. The message from this table is clear: the lack of sensitivity in Bob's average posteriors to information quality is predominantly driven by the lack of sensitivity in Bob's conditional posteriors. In fact, the model in which Anne uses Bayesian signal frequencies performs just as well as the one in which she augments signal frequencies through the lens of Grether's model.

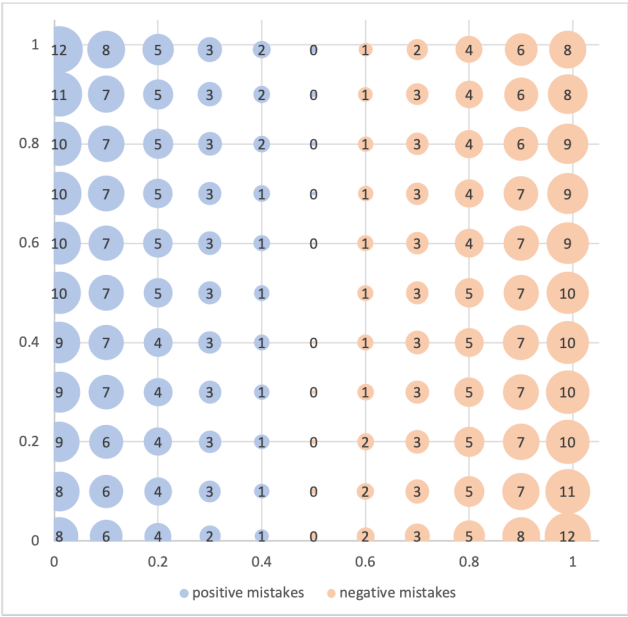
Observation 6: Grether's model offers a parsimonious explanation for the lack of responsiveness in average posteriors to information quality. This non-responsiveness is primarily driven by the lack of sensitivity in Bob's conditional posteriors to the information quality he is exposed to.

Magnitudes of Anne's Mistakes. How accurately does Anne predict Bob's expected posteriors? Figure 3.8 illustrates the differences between Bob's actual and Anne's predicted posteriors for signals with 90% accuracy using the estimated parameters of Grether's model reported in Table 3.4.[33] In the figure, positive mistakes are represented by blue circles, while negative mistakes are shown as orange circles. The size of each circle, along with the number inside it, indicates the magnitude of the mistake for a specific pair of Anne's and Bob's priors, with Anne's priors depicted on the horizontal axis and Bob's priors on the vertical axis. A positive

[33] We use parameters reported in the second-to-last row of Table 3.4. The figure depicts mistakes for the following values of priors: 0.01, 0.1, 0.2, ..., 0.9, and 0.99. A similar analysis for weak signals with 65% accuracy is provided in Figure C.8 in the Appendix, showing similar results.

mistake means Anne predicts Bob's expected posterior to be lower than it actually is. Conversely, a negative mistake indicates the opposite, i.e., Anne's prediction about Bob's expected posterior is higher than it is.

Figure 3.8: The difference between Bob's actual average posteriors and Anne's prediction of Bob's average posterior based on estimates of Grether's model, signal accuracy 90%



Notes: For each pair of Anne and Bob's priors, the size of the mistake is represented by the bubble size, with the exact value displayed inside the bubble. Anne's priors are shown on the horizontal axis, while Bob's priors are depicted on the vertical axis. Both priors take values of 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99. The mistake estimates are derived using the parameters of Grether's model reported in Table 3.4 for the dataset excluding corner beliefs. Blue bubbles represent positive mistakes, where Anne overestimates Bob's expected posterior (i.e., she believes Bob's posterior is higher than it actually is). In contrast, orange bubbles indicate negative mistakes, where Anne underestimates Bob's expected posterior (i.e., she believes it is lower than it actually is).

Figure 3.8 demonstrates that, apart from extreme priors, Anne is fairly accurate in predicting Bob's expected posteriors. For all priors except the most extreme cases (0.01 and 0.99), the prediction errors are at most 7 percentage points, which is remarkably good. The largest mistakes occur for highly polarized priors. For example, when Anne holds a prior of 0.01 and attempts to predict Bob's expected posterior given his prior of 0.99, she overestimates the extent to which Bob's belief shifts towards hers predicting that Bob's expected posterior will be 12 percentage points lower than it actually is. Similarly, when Anne's prior is 0.99 and she predicts

Bob's posterior given his prior of 0.01, Anne again overestimates the magnitude of Bob's shift toward her high prior by 12 percentage points, resulting in a negative mistake.

These errors are consistent with the estimates of Grether's model reported in Table 3.4. As we've shown, Anne believes that Bob places less weight on his own prior than he actually does, i.e., $c^{\text{Bob}} < c^{\text{Anne}}$. Consequently, when the two have very extreme and polarized priors, Anne thinks that Bob will on average shift further away from his prior relative to what he actually does. This highlights the limitations of Anne's predictions in such scenarios.

Observation 7: Anne is quite accurate in predicting movements in Bob's expected posteriors. However, the largest errors arise when Anne and Bob hold highly polarized and extreme priors. In these cases, Anne overestimates the extent to which Bob's beliefs shift toward her own due to information.

6 Conclusions

This paper provides empirical evidence on how people think others revise their beliefs in response to new information. Our findings show that individuals generally believe others' beliefs follow the Martingale property—i.e., from an ex-ante perspective, new information cannot systematically shift beliefs in one direction. However, we find only partial support for the Information Validates the Prior (IVP) property. Specifically, while people do expect new information to bring others' beliefs closer to their own effectively reducing polarization of opinions, the degree of this adjustment is less sensitive to information quality than predicted by the Bayesian model. This reduced sensitivity stems from flatter-than-expected conditional posteriors and signal frequencies. Moreover, we observe that even extreme or "corner" beliefs are not entirely degenerate, as individuals are open to revising them, and they believe others will do the same when confronted with contradictory evidence.

Our findings carry important implications for various strategic environments. The rigidity of others' beliefs and their limited responsiveness to information quality can be both advantageous and disadvantageous, depending on the setting. From a policy standpoint, a lack of responsiveness to high-quality information is often problematic, as information campaigns are designed to shift public beliefs, influence subsequent actions, and regulate markets. However, in certain scenarios, this reduced sensitivity may prove beneficial. To illustrate, consider the voluntary testing game, in which

an agent with private knowledge about their ability or product quality can choose to undergo a costly test that generates an independent public signal of quality. The agent's payoff is based on the market's posterior belief of their quality minus the cost of testing. Navin Kartik et al., 2021 theoretically demonstrate that, under standard informational assumptions, more informative tests lead to lower participation rates. However, our results suggest that participation will be less responsive to test quality, which, in this case, might be a welfare-improving outcome.

Our findings have also implications for information design literature. When people anticipate others to be relatively unresponsive to the quality of information, it may be more effective to expose them to a sequence of weak signals than a single strong signal, even if the collection of weak signals in theory conveys the same amount of information as a strong signal alone. We are hoping future research will provide empirical evidence on response to these types of information framings.

C APPENDIX FOR CHAPTER 3

In this section, we present additional data analysis, which is referenced in the paper.

Table C.1: Differences in Anne's and Bob's beliefs before and after Bob consumes new evidence when Anne and Bob have different priors, but the difference is at most 40 pp, in absolute terms.

Bob's prior is 5 to 40 pp different from Anne's prior

	Anne's prior							
	all		extreme		intermediate		close to uniform	
	mean (se)	med	mean (se)	med	mean (se)	med	mean (se)	med
before info	21.4 (0.4)	20	17.4 (0.6)	10	23.9 (0.7)	23.5	26.7 (1.2)	30
info acc 90%	18.8 (1.0)	15	13.8 (1.4)	10	22.8 (1.6)	20	22.7 (1.6)	25
info acc 65%	17.4 (0.9)	13	17.6 (1.3)	11.5	16.2 (1.4)	12	18.1 (1.9)	15
p-value								
before vs 90%	$p < 0.001$		$p < 0.001$		$p = 0.399$		$p = 0.045$	
before vs 65%	$p < 0.001$		$p = 0.537$		$p < 0.001$		$p < 0.001$	
90% vs 65%	$p = 0.175$		$p = 0.033$		$p = 0.001$		$p = 0.051$	

Notes: The table reports the difference between Anne's and Bob's prior beliefs in the first row and the differences between Anne's beliefs about Bob's expected posterior and her own prior in the second and third rows. The second and the third rows differ by signal accuracy. We focus exclusively on cases where Anne and Bob have different priors. Entries marked n/a indicate instances with fewer than 10 observations. Anne's prior is categorized into three groups: extreme priors (below 20 or above 80), close-to-uniform priors (between 40 and 60), and intermediate priors (those between 20 and 40 or between 60 and 80).

Figure C.1: Statements and Anne's Prior Beliefs (treatment T0, part 1)

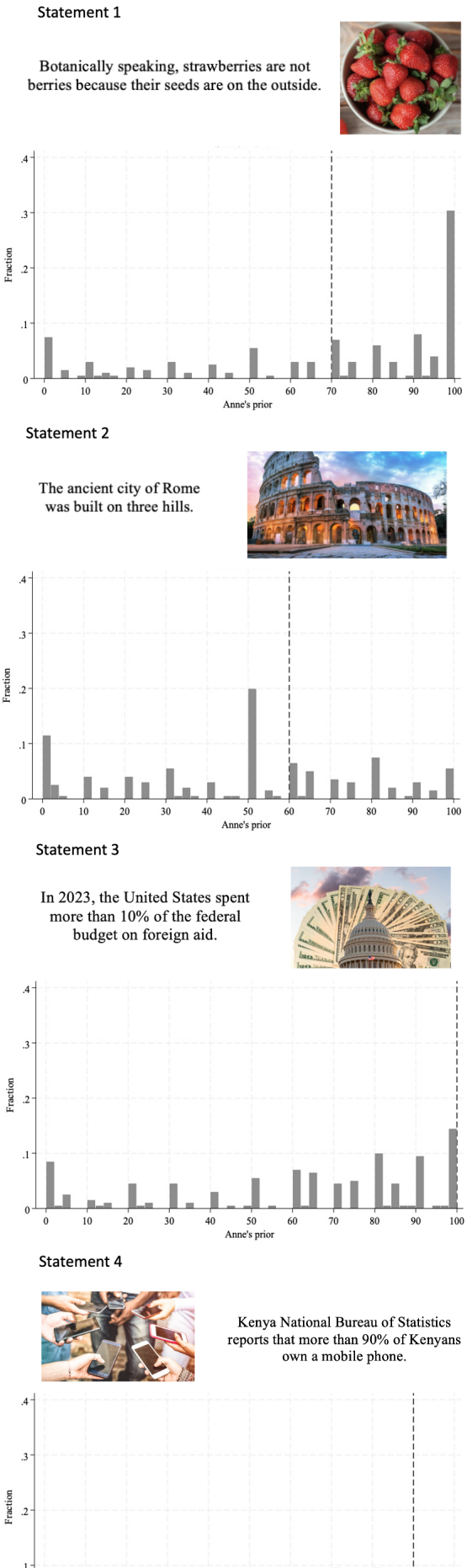
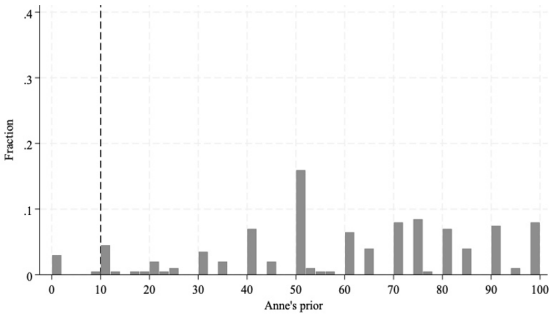


Figure C.2: Statements and Anne's Prior Beliefs (treatment T0, part 2)

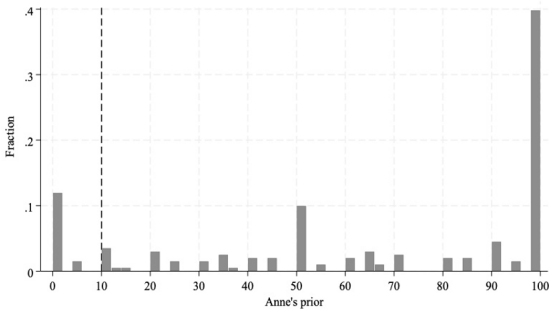
Statement 7

One cup of boiled broccoli contains more calcium than 10 dried figs.



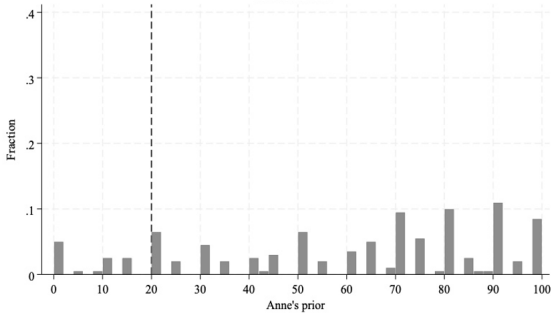
Statement 8

Pierre is the capital city of the U.S. state of South Dakota.



Statement 9

According to U.S. Bureau of Labor Statistics, the current unemployment rates in the U.S. are similar for both men and women, ranging between 3% and 4%.



Statement 10

According to the U.S. Census, in 2023, Black and African American residents comprised about 20% of the population in the United States.

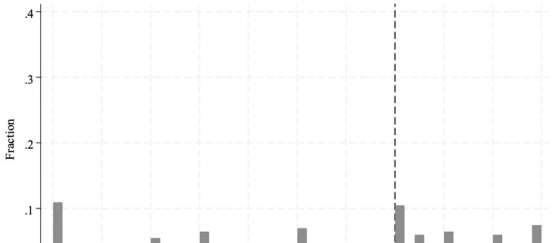
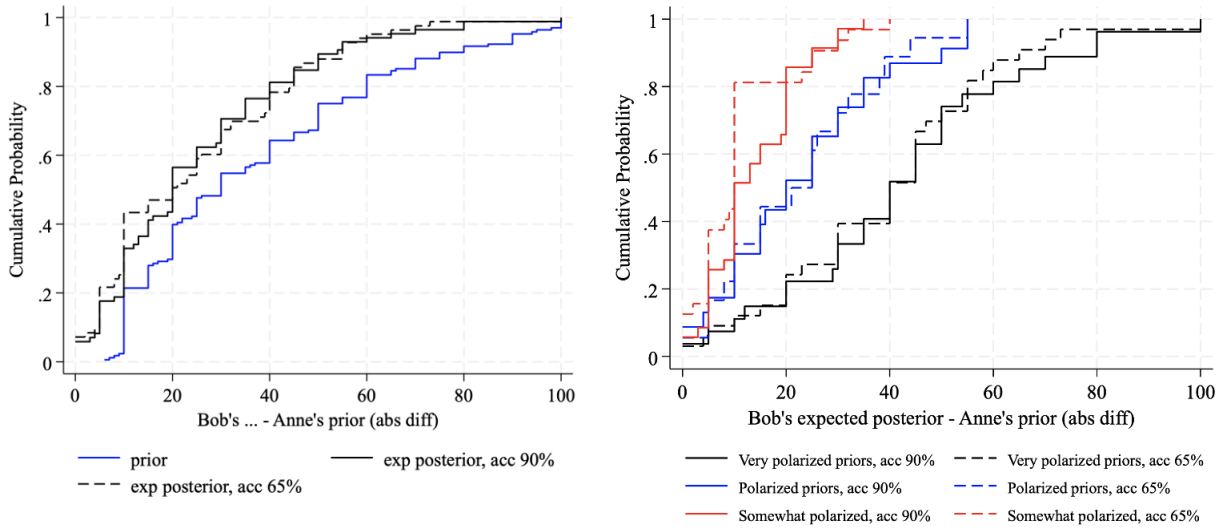
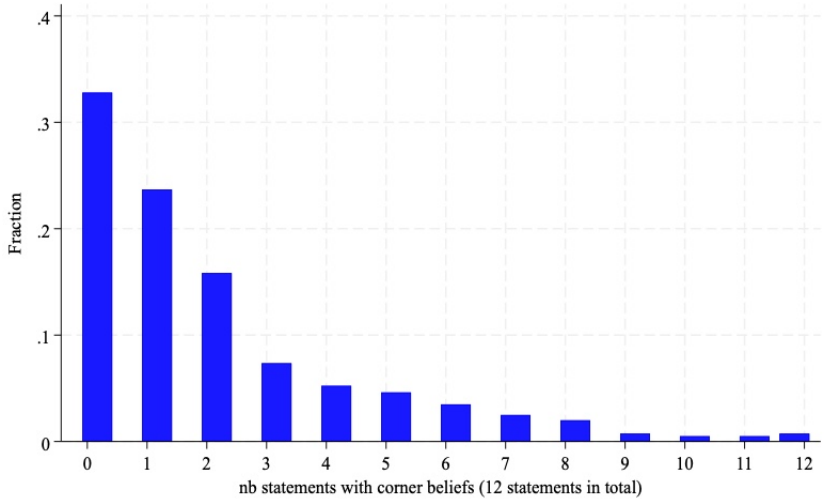


Figure C.3: Changes in Bob's beliefs when Anne and Bob have different priors for politically-charged statements (statements 3 and 6)



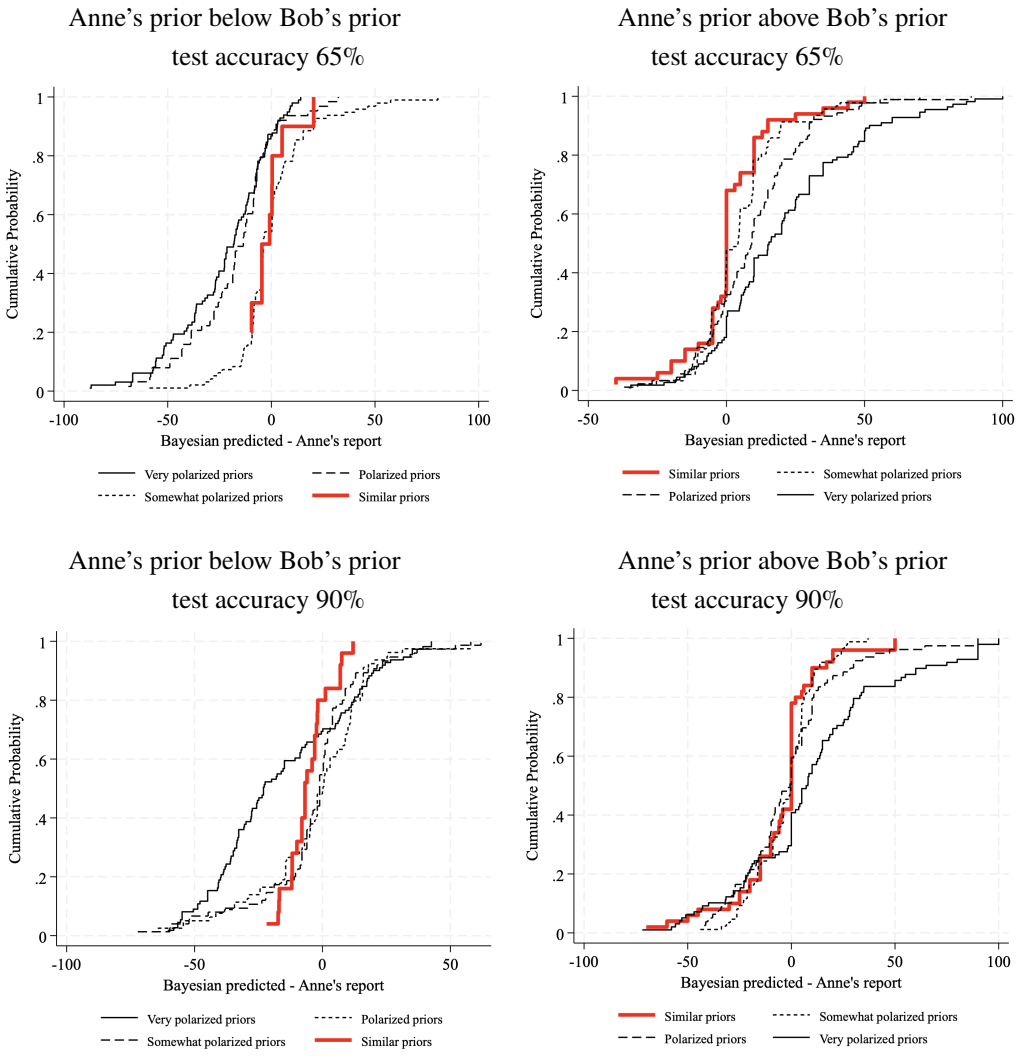
Notes: The left panel depicts the CDFs of the absolute differences between Bob's and Anne's priors, as well as the absolute differences between Bob's expected posteriors and Anne's priors. The right panel displays the differences between Bob's expected posteriors and Anne's priors, broken down by each level of prior disagreement. The analysis in both panels focuses on cases where Anne and Bob have different priors.

Figure C.4: How Often Anne Reports Corner Beliefs in All Rounds?



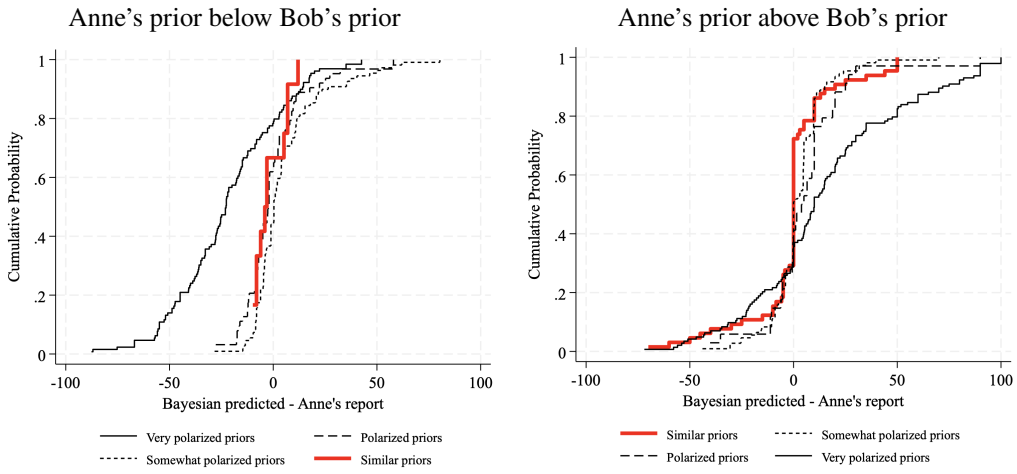
Notes: We present the histogram of the number of statements in which corner belief is reported at the individual level. Data is from both parts in treatment T0.

Figure C.5: Anne's estimates of Bob's expected posteriors vs Bayesian predictions, by information structure



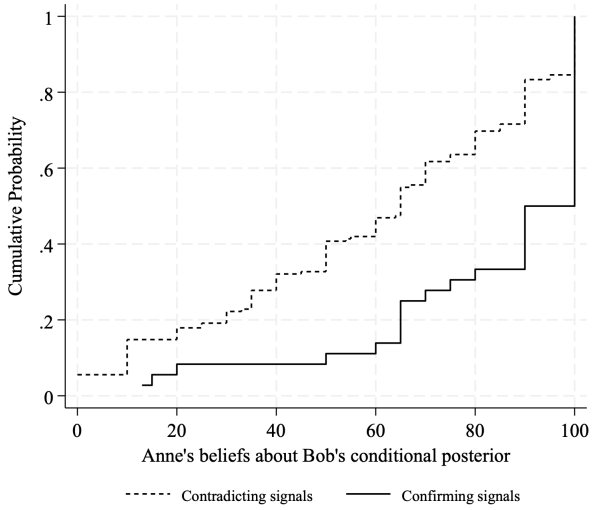
Notes: We plot the CDFs of the differences between Bob's Bayesian-predicted posterior expectations and Anne's estimates of these values, by signal accuracy. The top plots are for signal structure with accuracy 65% and the bottom plots are for the signal structure with accuracy 90%. The plots are separated into cases where Anne's prior is lower than Bob's (left panels) and cases where Anne's prior is higher than Bob's (right panels). The data is sourced from Part 2 of treatment T2.

Figure C.6: Anne's estimates of Bob's expected posteriors vs Bayesian predictions when Anne's priors are extreme



Notes: We plot the CDFs of the differences between Bob's Bayesian-predicted posterior expectations and Anne's estimates of these values. We focus on cases in which Anne's prior is extreme, i.e., below 20 or above 80. The plots are separated into cases where Anne's prior is lower than Bob's (left panel) and cases where Anne's prior is higher than Bob's (right panel). The data is sourced from Part 2 of treatment T2.

Figure C.7: How Anne Thinks Bob Updates his Corner Beliefs



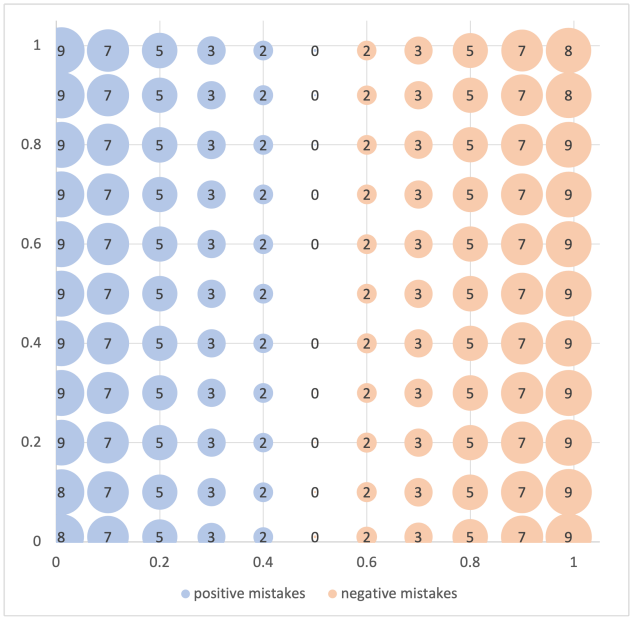
Notes: The figure depicts two CDFs, one for what Anne thinks Bob's posterior will be after observing a confirming signal and one for a contradicting signal. In both cases, Bob starts from the degenerate prior equal to 100. The data is from Part 2 of treatment T1.

Table C.2: Decomposition of the Combined Flattening Effect

	Posteriors cond on a signal	Signal frequency	Model Fit β_0	β_1	root MSE
(1)	Bayesian	Bayesian	0.27** (0.01)	0.57** (0.01)	0.2303
(2)	Bayesian	Grether	0.17** (0.01)	0.76** (0.02)	0.2346
(3)	Grether	Bayesian	0.06** (0.01)	0.97** (0.02)	0.2275
(4)	Grether	Grether	0.07** (0.01)	0.96** (0.02)	0.2288

Notes: We use the data from all treatments but exclude corner priors.

Figure C.8: The difference between Bob's actual average posteriors and Anne's prediction of Bob's average posterior based on estimates of Grether's model, signal accuracy 65%



Notes: For each pair of Anne and Bob's priors, the size of the mistake is represented by the bubble size, with the exact value displayed inside the bubble. Anne's priors are shown on the horizontal axis, while Bob's priors are depicted on the vertical axis. Both priors take values of 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99. The mistake estimates are derived using the parameters of Grether's model reported in Table 3.4 for the dataset excluding corner beliefs. Blue bubbles represent positive mistakes, where Anne overestimates Bob's expected posterior (i.e., she believes Bob's posterior is higher than it actually is). In contrast, orange bubbles indicate negative mistakes, where Anne underestimates Bob's expected posterior (i.e., she believes it is lower than it actually is).

C.1 Alternative Structural Models

In this section, we estimate beliefs' revisions through alternative structural models proposed in the literature and run the horse race between these models. We do this separately for Anne's own beliefs and Anne's beliefs about Bob's conditional posteriors. We consider four alternative models:

1. BAYESIAN model, according to which the posterior-odds ratio depends on signal precision θ and Anne's prior a_0, i.e.,

$$\frac{a_{s=1}}{1-a_{s=1}} = \frac{a_0}{1-a_0} \cdot \frac{\theta}{1-\theta}.$$

2. BASE-RATE NEGLECT (BRN) model, according to which Anne completely ignores her prior and, as a result, the posterior-odds ratio depends only on the signal-odds ratio, i.e.,

$$\frac{a_{s=1}}{1-a_{s=1}} = \frac{\theta}{1-\theta}.$$

3. COGNITIVE IMPRECISION model of Woodford, 2020, according to which Anne misperceives signal strength but otherwise uses the Bayes' rule. In particular, we follow Augenblick et al., 2024 paper and define the true signal-odds ratio as $\mathbb{S} = \log\left(\frac{\theta}{1-\theta}\right)$ and perceived signal-odds ratio as $\mathbb{E}\left(\hat{\mathbb{S}}\right) = k \cdot \mathbb{S}^{\beta}$. Then, the difference between posterior-odds and prior-odds ratios in log terms can be written as

$$\log\left(\frac{a_{s=1}}{1-a_{s=1}}\right) - \log\left(\frac{a_0}{1-a_0}\right) = \log(k) + \beta \cdot \log\left(\frac{\theta}{1-\theta}\right).$$

   Using this formulation, we can estimate the two parameters of this model (k, β).

4. GRETHER model used in the paper, according to which the posterior-odds ratio in log terms can be written as

$$\log\left(\frac{a_{s=1}}{1-a_{s=1}}\right) = d \cdot \log\left(\frac{\theta}{1-\theta}\right) + c \cdot \log\left(\frac{a_0}{1-a_0}\right)$$

   and we estimate the two parameters of this model, (c, d), which represent how Anne under-/over- infers from her own prior and from the new signal she receives.

To judge which behavioral model fits our data best, we run a simple linear regression of observed posteriors on the predicted ones, clustering observations at the individual level:

$$\text{Observed Posterior} = \text{const} + \text{intercept} \cdot \text{Predicted Posterior} + \epsilon.$$

The best fit is achieved when the estimated constant is close to zero, the estimated intercept is close to one, and the value of the mean-squared errors is small.

Table C.3 presents the results and shows that Grether's model emerges as a clear winner among the considered alternatives. This model captures most variation in Anne's own posteriors as well as Anne's beliefs about Bob's conditional posteriors and significantly improves the fit relative to the Bayesian model, the BRN model, and the cognitive imprecision model.

Table C.3: Comparing the fit of different behavioral models

	BRN	BAYESIAN	COGNITIVE IMPRECISION	GRETHER
Anne's own posteriors				
const	0.33** (0.01)	0.27** (0.01)	0.21** (0.01)	0.12** (0.01)
intercept	0.48** (0.02)	0.55** (0.02)	0.63** (0.02)	0.84** (0.02)
root MSE	0.2514	0.2239	0.2312	0.2217
Bob's conditional posteriors				
const	0.36** (0.02)	0.35** (0.02)	0.34** (0.02)	0.16** (0.03)
intercept	0.40** (0.04)	0.44** (0.03)	0.42** (0.03)	0.80** (0.05)
root MSE	0.2470	0.2341	0.2481	0.2320

Notes: For Anne's own posteriors, we use data from both parts in T0 and part 1 in T1 and T2. For Anne's beliefs about Bob's conditional posteriors, we use the data from Part 2 in T1. In all estimations, we exclude corner priors and corner posteriors. This is done to maintain with results reported in Table 3.3 and general comparability across models. This is because Grether's and Woodford's models involve logs of prior-odds and posterior-odds ratios and, as a result, are not defined for corner priors and posteriors.

As a final exercise, we take the cognitive imprecision model and the estimated parameters (k, β) obtained by Augenblick et al., 2024 and ask what would these estimates predict in our experiment. Augenblick et al., 2024 obtains $k = 0.88$ and $\beta = 0.76$ which imply very close to Bayesian posteriors for low-precision signals (65% accuracy) and significant underinference relative to Bayesian posteriors for high-precision signals (90% accuracy). These predictions do not fit our data as Figure 3.4 clearly shows.

Our discussion above supports the use of the Grether model for analyzing revisions of beliefs in a structural manner.

References

Agranov, M. [M.], Dasgupta, U., & Shotter, A. (2024). Trust me: Competition and communication in a psychological game. *Journal of the European Economic Association* (cit. on p. 103).

Agranov, M. [Marina], & Reshidi, P. (2024). Disentangling suboptimal updating: Task difficulty, structure, and sequencing. *working paper* (cit. on p. 102).

Alonso, R., & Camara, O. (2016). Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, *165*, 672–706 (cit. on p. 101).

Andreoni, J., & Mylovanov, T. (2012). Diverging opinions. *American Economic Journal: Microeconomics*, 209–232 (cit. on p. 110).

Augenblick, N., Lazarus, E., & Thaler, M. (2024). Overinference from weak signals and underinference from strong signals. *Quarterly Journal of Economics*, *forthcoming* (cit. on pp. 102, 140, 142).

Azzimonti, M., & Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, *76* (cit. on p. 100).

Ba, C., Bohren, A., & Imas, A. (2023). Over- and underreaction to information. *working paper* (cit. on p. 102).

Becker, G., DeGroot, M., & Marschak, J. (1964). Measuring utility by a single response sequential method. *Behavioral Science*, *9*, 226–232 (cit. on p. 109).

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations*, *1*, 69–186 (cit. on pp. 100, 102, 110, 117, 119, 123).

Bikhchandani, S., Hirshleifer, D., Tamuz, O., & Welch, I. (2024). Information cascades and social learning. *Journal of Economic Literature* (cit. on p. 97).

Calford, E., & Chakraborty, A. (2023). Higher-order beliefs in a sequential social dilemma. *Working Paper* (cit. on p. 103).

Carlsson, H., & van Damme, E. (1993). Global games and equilibrium selection. *Econometrica*, *61(5)*, 989–1018 (cit. on p. 97).

Charness, G., & Dufwenberg, M. (2006). Promises and partnerships. *Econometrica*, *74*, 1579–1601 (cit. on p. 103).

Charness, G., Gneezy, U., & Rasocha, V. (2014). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior and Organization*, *189*, 234–256 (cit. on p. 102).

Che, Y., & Kartik, N. [N.]. (2009). Opinions in incentives. *Journal of Political Economy*, *117*, 815–860 (cit. on p. 101).

Danz, D., Madarasz, K., & Wang, S. (2024). The biases of others: Projection equilibrium in an agency setting. *working paper* (cit. on pp. 104, 121).

Danz, D., Vesterlund, L., & Wilson, A. (2021). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, *112(9)*, 2851–2883 (cit. on p. 109).

DellaVigna, S., & Kaplan, E. (2007). The fox news effect: Media bias and voting. *Quarterly Journal of Economics*, *122(3)*, 1187–1234 (cit. on p. 100).

Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, *30*, 163–182 (cit. on p. 103).

Enke, B., & Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics* (cit. on pp. 102, 117, 127).

Esponda, I., Vespa, E., & Yuksel, S. (2023). Mental models and learning: The case of base-rate neglect. *American Economic Review* (cit. on pp. 102, 123).

Evdokimov, P., & Garfagnini, U. (2022). Higher-order learning. *Experimental Economics*, 1234–1266 (cit. on p. 103).

Fedyk, A. (2024). Asymmetric naivete: Beliefs about self-control. *Management Science*, *forthcoming* (cit. on p. 104).

Francetich, A., & Kreps, D. (2014). Bayesian inference does not lead you astray... on average. *Economics Letters*, *125*, 444–446 (cit. on p. 98).

Friedenberg, A., & Kneeland, T. [T.]. (2024). Beyond reasoning about rationality: Evidence of strategic reasoning. *working paper* (cit. on p. 103).

Garrett, R. (2009). Echo chambers online? politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, *14(2)*, 265–285 (cit. on p. 101).

Gneezy, U., Enke, B., Hall, B., Martin, D., Nelidov, V., Offerman, T., & van de Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics Studies* (cit. on pp. 102, 123).

Grether, D. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics*, *95*, 537–557 (cit. on pp. 100, 117, 124).

Healy, P. J. (2024). Epistemic experiments: Utilities, beliefs, and irrational play. *working paper* (cit. on p. 102).

Healy, P., & Leo, G. (2024). Belief elicitation: A user's guide. *Handbook of Experimental Economics Methodology* (cit. on p. 102).

Hirsch, A. (2016). Experimentation and persuasion in political organizations. *American Political Science Review*, *110*, 68–84 (cit. on p. 101).

Kartik, N. [Navin], Lee, F. X., & Suen, W. (2021). Information validates the prior: A theorem on bayesian updating and applications. *American Economic Review: Insights*, *3(2)*, 165–182 (cit. on pp. 98, 101, 105, 132).

Kneeland, T. [Terri]. (2015). Identifying higher-order rationality. *Econometrica*, *83*, 2065–2079 (cit. on p. 103).

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2002). Projection bias in predicting future utility. *Quarterly Journal of Economics* (cit. on p. 121).

Madarasz, K. (2016). Projection equilibrium: Definition and applications to social investment, communication and trade. *working paper* (cit. on p. 104).

Manski, C., & Neri, C. (2013). First- and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior*, *81*, 232–254 (cit. on p. 102).

Martin, G., & Yurukoglu, A. (2017). Bias in cable news: Persuasion and polarization. *American Economic Review*, *107*, 2565–2599 (cit. on p. 100).

McCarthy, N. (2019). Polarization: What everyone needs to know. *Oxford University Press* (cit. on p. 100).

McCarthy, N., Poole, K., & Rosenthal, H. (2006). Polarized america: The dance of ideology and unequal riches. *MIT Press* (cit. on p. 100).

McGranaghan, C., O'Donoghue, T., Nielsen, K., Somerville, J., & Sprenger, C. (2024). Distinguishing common ratio preferences from common ratio effects using paired valuation tasks. *American Economic Review* (cit. on p. 107).

Morris, S., & Shin, S. (2002). Social value of public information. *American Economic Review*, *92(5)*, 1521–1534 (cit. on p. 97).

Schlag, K., Tremewan, J., & van der Weele, J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, *18*, 457–490 (cit. on p. 102).

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87(3)*, 355–374 (cit. on p. 97).

Stroud, N. (2010). Polarization and partisan selective exposure. *Journal of Communication*, *60(3)*, 556–576 (cit. on p. 101).

Szkup, M., & Trevino, I. (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, *124*, 534–553 (cit. on p. 103).

Thaler, M. (2024). The fake news effect: Experimentally identifying motivated reasoning using trust in news. *American Economic Journal: Microeconomics*, 1–38 (cit. on pp. 102, 110).

Trevino, I., & Schotter, A. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, *6*, 103–128 (cit. on p. 102).

Trujano-Ochoa, D. (2024). Do others learn like me? higher order willingness to pay for information. *working paper* (cit. on pp. 103, 104).

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601 (cit. on pp. 100, 140).

Yuksel, S., & Oprea, R. (2022). Social exchange of motivated beliefs. *Journal of the European Economic Association*, 667–699 (cit. on pp. 100, 125).