Essays in Behavioral Economics

Inauguraldissertation
zur Erlangung des Grades eines Doktors
der Wirtschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
MARTA MAŁGORZATA KOZAKIEWICZ
aus Kozienice, Polen

2024

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Lorenz Götte
Zweitreferent:	Prof. Dr. Florian Zimmermann

Tag der mündlichen Prüfung: 30.09.2024

Contents

List of Figures

List of Tables

Introduction

This dissertation explores the question of how people learn if they desire to hold certain beliefs. When deciding whether to incorporate new information into pre-existing knowledge, an agent faces a trade-off between belief accuracy and its desirability (Bénabou and Tirole, 2016).[1] Studying the formation of beliefs in the presence of conflicting motives is crucial for understanding various behavioral phenomena, such as overconfidence, the demand for information, and belief polarization. These phenomena have important implications not only for individual decision-makers but also for various types of organizations and society as a whole. For example, both individuals and private companies can be affected by the negative consequences of overconfidence, e.g., excessive selection into competitive environments (Camerer and Lovallo, 1999; Niederle and Vesterlund, 2007), excessive trading (Barber and Odean, 2001), or suboptimal investment decisions (Malmendier and Tate, 2005, 2008). Moreover, overconfidence and belief polarization can impact democratic systems and the functioning of modern societies—polarization of political beliefs being the prime example (McCarty et al., 2016). Overconfidence is an important predictor of ideological extremeness, voter turnout, and partisan identification (Ortoleva and Snowberg, 2015). Yet, there are still open questions regarding how overconfidence arises, why it persists, and what can be done to prevent it. This dissertation includes three self-contained essays, each considering a different question on the process of belief formation that can leave the decision-maker with biased beliefs. In the first essay, I demonstrate how the learning process of an overconfident agent can go awry, making the agent more mistaken about the state of the world with every decision he makes. In the second essay, I take a closer look at the fundamental processes that govern belief formation, as I examine how people interpret favorable and unfavorable feedback. The final essay describes the complexities of belief formation, uncovers the underlying factors, and links them to overconfidence. The three essays constitute the chapters of this thesis. I describe them in more detail below.

[1] For example, we all want to see ourselves as smart, attractive, moral individuals with a bright future ahead of us. Learning about our traits or future prospects is likely to be affected by our preferences.

In the first chapter, *"Experimental Evidence on Misguided Learning"* (joint work with Lorenz Götte), we study how people form beliefs in environments with multiple unknown parameters, some of which are relevant to agents' self-esteem. In particular, we examine how initial bias in beliefs about an ego-relevant characteristic affects learning about the state of the world. Using data from a laboratory experiment, we demonstrate that the learning process of an overconfident agent is *self-defeating*: the agent repeatedly takes suboptimal actions, misinterprets the output, and forms increasingly mistaken beliefs about the state. Therefore, we corroborate the theory of misguided learning formulated by Heidhues et al. (2018). We provide the first empirical evidence that allowing a biased agent to experiment and acquire new information is not only ineffective but in some cases counterproductive. Furthermore, we move beyond the theory as we examine how learning about multiple parameters evolves in ego-relevant and ego-neutral environments. When none of the parameters point to their characteristics, subjects are engaging in self-defeating learning to a lesser extent. This is partly because overconfident subjects are more willing to revise their beliefs about a neutral parameter downwards. The results show that misguided learning is more likely to arise (and persist) when one's ego is at stake.

In the second chapter, *"Belief-based Utility and Signal Interpretation"*, I attempt to answer the following question: Do people perceive favorable feedback in a different way than unfavorable one? The existing literature disagrees not only on the magnitude but also the direction of the bias (Benjamin, 2019). Using data from a new experiment, I provide strong evidence that people perceive favorable feedback as more likely to be informative. Furthermore, I design a new control condition to better understand the nature of the bias. Participants in the control group evaluated the informativeness of a signal ex-ante, conditioned on possible signal realizations. By comparing beliefs reported after a signal to the reports stated ex-ante, I show that subjects distort their perception in a motivated way *after* receiving a signal. The effect of receiving a signal amounts to 10.6 percentage points (a 27.9% increase in relative terms) for favorable feedback. In other words, when a person gets a signal that he is smart, he becomes 27.9% more certain that this signal is accurate, compared to the evaluation of the same signal ex-ante. There is no significant difference after unfavorable feedback. Consequently, even though signals significantly shifted subjects' beliefs, they did it selectively, and the average overconfidence level remained the same at the end of the study. The results cast a

new light on the origins of overconfidence, pointing towards the role of affect (or utility from beliefs shifted by the signal) in asymmetric updating.

In the final chapter, *"Hope for the Best, Prepare for the Worst: Signal Anticipation and Ex-ante Belief Manipulation"*, I further investigate how people form beliefs when they derive utility from their self-perception. In this part, I experimentally test a model of belief choice with reference-dependent utility. The basic idea is that people can "prepare themselves" for the arrival of new information by adopting overly pessimistic beliefs. By distorting her prior beliefs, an agent can 1) hedge against a painful downward shift in beliefs after a negative signal and 2) enhance a pleasant surprise from a positive signal. To test the model, I designed a lab experiment in which subjects solve an IQ test and subsequently report beliefs about their relative performance. I introduce an exogenous variation in subjects' expectations over the upcoming signal, which allows me to identify belief manipulation. The results confirm the main predictions of the model, substantiating the claim that utility from beliefs is reference-dependent. Moreover, I examine a previously unexplored link between gain-loss attitudes and overconfidence and confirm it in the data. As predicted, overconfident subjects tend to be low-ability and non-loss-averse, and they end up with a bias that is larger than the bias of underconfident participants. These results provide further support for the theory and bring us one step closer to understanding the sources of overconfidence.

The conclusions of the three chapters suggest that, although the process of belief formation is rather complex, we can identify its systematic components. Learning about ego-relevant parameters is driven by two distinct forces: the need for accuracy and the desire to maintain a positive self-view. The latter induces the agent to interpret feedback in a self-serving manner by misattributing it to the state of the world (as described in Chapter 1) or by distorting his beliefs about signal informativeness (discussed in Chapter 2). However, the hedonic motive does not always direct the agent towards a more optimistic belief. As shown in Chapter 3, the decision-maker might adopt pessimistic beliefs to make himself feel good in the future—after receiving new information. In this case, the extent of belief manipulation is determined by individual parameters such as the agent's true ability and his aversion to losses. Investigating the interplay between these factors and motives gives us a fuller picture of how people learn about themselves and why they end up with biased beliefs.

CHAPTER 1

Experimental Evidence on Misguided Learning

Many economic decisions require an accurate assessment of the state of the world. Often, more than one decision-relevant aspect is unobservable, and people have to form beliefs *simultaneously* about multiple parameters. Learning in such environments is particularly challenging. Agents need to not only keep track of actions and their consequences but also disentangle the effects of various factors in order to update beliefs about specific parameters. If an agent holds biased beliefs about one parameter, e.g. overconfident beliefs about his own ability, he is likely to make incorrect inferences from the observed data.

Heidhues et al. (2018) show that, in some cases, the learning process goes awry: the agent repeatedly misinterprets the data, takes suboptimal actions, and forms more and more incorrect beliefs about the state of the world.[1] Learning is "misguided" and, since it is the agent who generates the observations that lead him astray, one can describe it as "self-defeating". Importantly, even if initially the agent has correct beliefs about *all* other aspects of the world, biased perception of ability can start a process that drives beliefs away from the truth. The theory predicts this pattern in a range of applications: it can explain why overconfident individuals exert too little effort, managers do not delegate enough tasks to subordinates, and CEOs engage in unprofitable mergers. It can also explain why additional feedback doesn't always help to correct one's actions. Yet, the extent to which people engage in misguided learning has not been examined.

In this paper, we use data from a carefully designed laboratory experiment to provide the first empirical evidence on misguided learning.[2] We test the comparative statics of the model

[1] An illustrative example considers an overconfident agent who is learning about the state of the world by taking actions and observing output, which also depends on his unknown ability. After observing an unexpectedly low output, the agent does not interpret it as a result of his low ability but concludes that the state must be worse than expected. He adjusts his action to match the new belief about the state. The increased mismatch between the action and the state further lowers the output. This reinforces the agent's pessimism and leads to an action that, in reality, fits even worse. Over time, the agent takes more and more inadequate actions and becomes increasingly mistaken about the state.

[2] A laboratory setting is particularly suitable for studying misguided learning. Firstly, it enables us to elicit beliefs multiple times in an incentive-compatible way, providing a precise measure of overconfidence and changes in beliefs (rarely observable in the field). Secondly, it gives us tight control over technology and information available

by Heidhues et al. (2018) and document the learning processes of biased agents. Secondly, we move beyond the model as we investigate how learning about multiple parameters evolves in ego-relevant and ego-neutral environments.

Our experiment integrates all features of the model in a simple way. The main goal was to create an environment in which subjects take actions, observe output and learn about the underlying state of the world. Importantly, the output also depends on an unknown parameter that is relevant to subjects' self-esteem. For this purpose, we used subjects' relative performance in an IQ test taken in the first part of the study.[3] Before the main task, we also elicited participants' beliefs about their relative performance.

In the second part of the experiment, participants completed several rounds of a learning exercise. In every round, the task was to estimate an unknown state of the world: a randomly drawn integer between -10 and 10. Participants had 4 trials to guess the state and were remunerated based on the accuracy of their guesses. After making a guess, each participant received feedback in the form of a real number between 4 and 51 displayed on the individual computer screen. Feedback was determined by the state of the world and one's relative performance in the IQ test. In every trial, the optimal strategy was to enter one's best guess about the state of the world. Therefore, we could directly track participants' belief formation process. After the learning exercise, we again elicited subjects' beliefs about their relative performance.

To help participants correctly interpret the feedback, we provided them with tables to look up which states of the world and relative performances are consistent with the feedback they observed. We did not preclude subjects from considering different performance levels and they were free to choose any combination of the two parameters. Giving subjects the opportunity to reconsider their beliefs about ability allows us to see if misguided learning emerges even in environments where beliefs are not restricted.

We introduced two experimental conditions: treatment and control. In the treatment condition, participants received informative feedback based on their last guess, whereas in

to subjects. In the field setting, technology is usually unobservable, and without the knowledge of the output generating process, one cannot formulate the model's predictions.

[3]We decided to use intelligence as an input to the production function for several reasons. Firstly, intelligence is known as a personal characteristic that people deeply care about, so a measure of IQ seems to be a good candidate for a genuinely ego-relevant parameter. Secondly, the literature provides evidence that people have biased assessments of their relative cognitive ability (see, for example, Burks et al., 2013). One would expect misguided learning to arise in this context. Last but not least, cognitive ability as a component of human capital is an actual input to many production functions.

the control condition, feedback did not depend on subjects' guesses. Thus, the main mechanism of the model – the interdependence between actions and feedback – was shut down in the control condition. We kept all other features of the experiment unchanged: participants were asked to make four guesses and after each one, a number was displayed on their computer screens. We use the control condition to exclude alternative explanations, for example, that the effect is an artifact of repeated choice-based elicitation. Rather, we show that it is induced by informative feedback as in Heidhues et al. (2018). The experiment was conducted in November 2017 in BonnEconLab at the University of Bonn. We collected data from 171 subjects, mostly university students.

Furthermore, we designed an ego-neutral condition in which output depends on a parameter that does not affect agents' self-esteem. In this condition, subjects performed the task based on the performance of some other, randomly selected individual who reported similar beliefs (we assume that the performance of another subject is irrelevant to one's ego).[4] The structure of the experiment was identical to that of the main condition, and it allows us to isolate the effect of ego-relevance of one of the parameters. The data from 155 participants in the ego-neutral condition was collected in November 2018.

Overall, we find strong support for the predictions of the misguided learning model. When overconfident individuals can adjust their actions and learn about the state of the world, repeated feedback leads them to form increasingly mistaken beliefs. Their learning process is self-defeating: overconfident participants tend to attribute an unsatisfactory outcome to the realized state instead of their relative performance, take suboptimal actions, and become pessimistic about the state over time. Importantly, we find a significant difference between the treatment and the control condition, confirming that the effect is driven by the mechanism described in Heidhues et al. (2018).

The effect is more pronounced for participants who are more biased about their ability. We test the model's comparative statics and show that the more overconfident the participant is, the more mistaken about the state he becomes. We also corroborate the qualitative predictions for underconfident and unbiased subjects.[5] The learning process of the underconfident

[4]After eliciting beliefs about subjects' relative performance, participants were informed that they will be randomly matched to a person from one of the previous sessions, who took the same IQ test and reported the same beliefs but not necessarily obtained the same IQ test score. Before the main task, we elicited subjects' beliefs about the relative performance of the person matched to them.

[5]We use the term "misguided learning" to describe the learning of biased (over- or underconfident) agents, and we only refer to the overconfident agents' misguided learning as "self-defeating learning".

agent is misdirected and its trajectory is different from that of the overconfident agent. We do not detect any similar pattern in the behavior of unbiased participants. In line with the model, the unbiased subjects immediately learn the true state and take the optimal action in the following trials.

However, the effects observed in the data are quantitatively lower than the theory predicts. The gap between the theoretical predictions and the observed behavior is caused by some participants updating their beliefs about ability during the experiment. We observe a significant difference in beliefs measured before and after the learning exercise, with 25% of participants revealing unbiased beliefs after the task (compared to 7.6% before the task). Notwithstanding, almost 80% of subjects who were classified as overconfident at the beginning of the experiment remained overconfident, and many of them were engaging in self-defeating learning until the end of the last round.

Using data from the ego-neutral condition, we show that self-defeating learning is more likely to arise and persist when one's ego is at stake. When the output is based on the IQ test performance of some other, randomly selected individual, misdirected learning of overconfident agents is significantly mitigated. Overconfident participants in the ego-neutral condition are engaging in self-defeating learning to a lesser extent partly because they are willing to revise downwards their beliefs about the ability of their match.[6]

The opposite is true for underconfident agents. Underconfident participants are *more* likely to become unbiased in the ego-relevant condition, that is, when learning about their own ability, compared to similarly underconfident subjects in the ego-neutral condition. The results demonstrate that, when learning involves multiple parameters and some of them are ego-relevant, people are steered to learn along the dimension that brings them higher ego utility. While motivated attribution of ego-relevant outcome is a phenomenon well-known in the psychological literature (see Coutts et al., 2020, for a review of the literature), our paper is the first to demonstrate it in a dynamic setting.

Our work is partially motivated by the behavioral literature on motivated reasoning, which suggests that people might interpret feedback in a self-serving manner (see Bénabou and Tirole, 2016, for a comprehensive literature review). A large body of work demonstrates that

[6]Importantly, we are comparing overconfident subjects in the ego-relevant condition with similarly overconfident participants in the ego-neutral control. This allows us to control for any confounding factors and disentangle the effect of agents' bias from the ego-relevance of the unknown parameter.

people use various strategies to manipulate their beliefs to maintain a positive self-view. These strategies include information avoidance (Golman et al., 2017), selective recall (Chew et al., 2020; Zimmermann, 2020; Huffman et al., 2022), and asymmetric updating (Eil and Rao, 2011; Buser et al., 2018; Coutts, 2019; Möbius et al., 2022). Our experiment was not designed to test any of these mechanisms directly, but to examine how motivated reasoning unravels in a more complex, dynamic environment.

We view our paper as complementary to the literature investigating the consequences of holding inaccurate beliefs with some degree of persistence. Overconfidence, a widely-studied phenomenon by both psychologists and economists, is believed to generate great costs for both the individual and the society.[7] We contribute to the literature concerning the implications of overconfidence as we document its detrimental effect on learning.

Our work is based on a model by Heidhues et al. (2018) that we describe in Section 1.1. Learning with biased beliefs about ability was also studied by Hestermann and Le Yaouanq (2021). They also consider two parameters that are not separately identifiable, but in their model, the agent is learning about both.[8] The remaining literature on belief formation and learning focuses on failures of reasoning that are conceptually different from the one we study. One should mention the work on selective attention in learning (Schwartzstein, 2014, Hanna et al., 2014), redundancy neglect in social learning (Eyster and Rabin, 2014, Enke and Zimmermann, 2017), difficulties in hypothetical thinking (Charness and Levin, 2009, Esponda and Vespa, 2014, Esponda and Vespa, 2016), overlooking selection problems (Esponda and Vespa, 2018, Enke, 2020), and misattribution of reference dependence in learning from experience (Gagnon-Bartsch and Bushong, 2022, Bushong and Gagnon-Bartsch, 2023). Perhaps the closest to our work, Coutts et al. (2020) test two different theories of self-attribution bias and show that, although people tend to update more favorably about themselves than about their teammates, they do not attribute the negative outcome to the other player. We contribute to the literature by providing empirical evidence on misguided learning and taking the first step towards understanding how people learn in environments with multiple unknown parameters.

[7] Negative consequences of overconfidence include excessive selection into competitive environments (Camerer and Lovallo, 1999; Niederle and Vesterlund, 2007), excessive trading (Barber and Odean, 2001), suboptimal investment decisions (Malmendier and Tate, 2005, 2008), and political polarization (Ortoleva and Snowberg, 2015).
[8] Unfortunately, their framework is different from our experimental setup, and we cannot directly test the model's predictions.

The paper is organized as follows. In Section 1.1, we describe a simplified version of the model and its testable predictions. Section 1.2 outlines our experimental design and Section 1.3 presents the empirical results. In Section 1.4, we discuss the results of the ego-neutral condition. Section 1.5 concludes.

1.1. Theoretical Framework

In this section, we present a version of the misguided learning model by Heidhues et al. (2018) and state its testable predictions. We adopted a simplified version of the model in order to focus on testing the main mechanism.[9] For the general framework, as well as the proofs, we refer the reader to the original paper.

1.1.1. The Model of Heidhues et al. (2018). In each period $t \in \{1,2,3,...\}$, the agent produces an observable output q_t according to the following production function:

$$q_t = Q(e_t, A, \Phi) = A + \Phi - L(e_t - \Phi),$$

where $e_t \in (\underline{e}, \bar{e})$ denotes the agent's action in period t, $A \in \mathbb{R}$ is the agent's true ability, $\Phi \in (\underline{\phi}, \bar{\phi})$ is the unknown state of the world, and $L(\cdot)$ is a symmetric loss function with $L(0) = 0$ and $|L'(x)| < k < 1$ for all x. The loss is minimized when the agent matches his action to the state of the world. The state Φ is drawn from the continuous prior distribution $\pi_0 : (\underline{\phi}, \bar{\phi}) \to R_{>0}$, and the agent has a prior belief about the state $\phi_0 = 0$.

In each period, the agent takes an optimal action given his belief ϕ about the state Φ. To minimize the loss function, he chooses $e^*(\phi) = \phi$. The agent follows a myopic decision rule: the action maximizes the expected output in a given period.[10,11] In the first period, the optimal action is equal to the agent's prior belief: $e_1^* = \phi_0 = 0$. It produces the following output (normalizing $A = \Phi = 0$):

$$q_1 = Q(e_1, A, \Phi) = -L(0) = 0.$$

[9] In Heidhues et al. (2018), the agent observes multiple noisy output realizations in every period and updates his beliefs based on *the average* output in that period (he averages out the random component). We decided not to include this feature of the model, as we were concerned that biases in information aggregation could obscure the results.

[10] The assumption implies that there is no learning motive at play. The agent is neither intentionally experimenting nor gathering data about his environment to make better choices in the future. Misguided learning is a by-product of a sequential, short-sighted optimization.

[11] We decided not to impose this assumption onto participants in our experiment. However, we expected that the task will induce short-sighted behavior to some extent.

The agent observes the output q_1 and compares it to the output that he expected. The difference between the observed and the expected output depends on the direction and magnitude of the agent's bias.

1.1.2. Overconfidence and Self-Defeating Learning. An overconfident agent believes that his ability is $\tilde{a} > A$ (it is higher than his actual ability A). After taking an action $e_1^* = \phi_0 = 0$, he expects to observe the output $\tilde{q}_1$:

$$\tilde{q}_1 = Q(e_1, \tilde{a}, \phi_0) = \tilde{a} > 0.$$

The agent is not suffering from any other information-processing bias and uses Bayes' rule to update his beliefs about the state of the world. As in Heidhues et al. (2018), we assume that the agent never updates his beliefs about his ability (we discuss this assumption in Section 1.1.4). Consequently, he attributes the difference between q_1 and $\tilde{q}_1$ to the state of the world. The agent concludes that the state is *worse* than he thought and he adopts a new belief that is lower than his prior: $\phi_1 < \phi_0$.

In Period 2 the agent chooses $e_2^* = \phi_1$. He observes the output $-L(\phi_1)$, while he expected to produce $\tilde{a} > -L(\phi_1)$. Once again, he is surprised by the output and attributes the difference to the state of the world. As a result, he becomes even more pessimistic about the state:

$$\phi_2 < \phi_1 < \phi_0. \tag{1.1}$$

With each observation, the agent's beliefs are driven further away from the true state. Hypothesis 1.OC summarizes the learning process of an overconfident agent:

Hypothesis 1.OC (Overconfident Agents)
The learning process of an overconfident agent is self-defeating: in each period, after taking an action and observing the output, the agent forms increasingly pessimistic beliefs about the state.

The change in beliefs in each each period depends on the difference between q_1 and $\tilde{q}_1$, which is in turn a function of agent's bias $|\tilde{a} - A|$. An overconfident agent with a larger bias has higher output expectations relative to a less biased individual. He will be more surprised

by the actual output and will attribute this larger difference to the state of the world. As a result, he will form more biased beliefs about the state compared to a less overconfident agent.

Hypothesis 2.OC (Overconfident Agents)
An overconfident agent with a larger bias will form more pessimistic beliefs compared to a less overconfident agent, and will end up further away from the true state.

Under the model's assumptions, the agent's belief about the state is not decreasing indefinitely but converges to a unique limiting belief ϕ_∞. This limiting belief is stable in the sense that the agent has no incentive to abandon it – at this point, he ends the learning process. Intuitively, a stable belief is a point belief that induces action and output that exactly matches the agent's expectations, thereby confirming his belief. It could be found by setting the difference between the actual and the expected outputs to zero: $Q(e^*(\phi_\infty), A, \Phi) - Q(e^*(\phi_\infty), \tilde{a}, \phi_\infty) = 0$. With the loss-function specification, that condition reads:

$$(A - \tilde{a}) + (\Phi - \phi_\infty) - L(\Phi - \phi_\infty) = 0. \tag{1.2}$$

By rearranging the above equation one can derive a formula for the stable belief ϕ_∞. It is worth noting that the stable belief is a function of the agent's bias.

1.1.3. Underconfident and Unbiased Agents. The model also predicts the behavior of underconfident agents. The analysis is analogous, with the only difference that the agent underestimates his true ability, i.e. $\tilde{a} - A < 0$. With the normalization of $A = 0$, this implies $\tilde{a} < 0$. In Period 1, the agent chooses $e_1^* = \phi_0 = 0$. He observes the output of $-L(0) = 0$, while he expected to produce $\tilde{a} < 0$. The agent does not update his beliefs about his ability, but instead he looks for ϕ that would explain the output. The updated belief ϕ_1 is *higher* than his prior – the agent concludes that the state of the world is *better* than expected.

In Period 2, the agent chooses $e_2^* = \phi_1$. He observes the output of $-L(\phi_1)$, while he expected to produce $Q(e_2, \tilde{a}, \phi_1) = \tilde{a} + \phi_1$. The output falls short of his expectations, so he concludes that the state is *worse* than he thought. The adjustment in the following period goes in the right direction, bringing the agent closer to the true state. In contrast to the overconfident agent, the underconfident agent's misguided learning is *self-correcting*. The model predicts

that the underconfident agent's beliefs satisfy:

$$\phi_1 > \phi_0 \quad \wedge \quad \phi_2 < \phi_1. \tag{1.3}$$

This allows us to formulate the following prediction about the belief-formation process of underconfident agents:

Hypothesis 1.UC (Underconfident Agents)
The learning process of an underconfident agent is self-correcting: after the initial overly optimistic assessment, the agent corrects his beliefs downwards.

In the first period, an underconfident agent with a larger bias is more surprised by the output than an underconfident agent with a smaller bias. Because the agent attributes the entire difference to the state of the world, he becomes more mistaken about the state compared to the less biased individual.

Hypothesis 2.UC (Underconfident Agents)
In the first period, an underconfident agent with a larger bias forms beliefs that are further away from the true state compared to the beliefs of a less underconfident individual.

While we cannot form a hypothesis similar to Hypothesis 2 for underconfident agents in every period, one can use (1.2) to derive a testable prediction for long-term learning outcomes of overconfident and underconfident agents.[12]

Hypothesis 3.UC&OC (Stable Belief)
The stable belief of an underconfident (overconfident) agent with a larger bias lies further from the true state than the stable belief of a less underconfident (overconfident) agent.

Last but not least, the model characterizes the behavior of unbiased agents. An unbiased individual correctly evaluates his ability $\tilde{a} = A$. After choosing the optimal action in the first

[12]While we admit that our setting is more suitable to test the short-term dynamics of the model, we argue that we can provide some evidence on the long-term. We discuss this point in more detail in Section 1.3.

period, $e_1^* = \phi_0 = 0$, he observes exactly the output he expects: $\tilde{a} + \phi = 0 = -L(0)$. The unbiased agent has no reason to update his beliefs any further, implying:

$$\phi_2 = \phi_1 = \phi_0. \tag{1.4}$$

We summarize it in the following hypothesis:

Hypothesis 1.UB (Unbiased Agents)
The learning process of an unbiased agent is immediate and stable afterwards.

1.1.4. Quantitative Predictions. It is important to note that the model is based on the assumption that agents never update their beliefs about their ability. Although there is some evidence that people are reluctant to update beliefs about ego-relevant characteristics, especially if prompted to revise them downwards (see, for example, Eil and Rao, 2011), the assumption of no updating is rather extreme. Still, the qualitative predictions of the model will hold on aggregate if beliefs about ability are sufficiently sticky. Even if some agents correctly update their beliefs, as long as the bias is not entirely reduced in the population we will observe misguided learning. In this case, one would expect the effect to be of the same direction, but quantitatively lower than predicted by the model.

1.2. Experimental Procedures

The experiment took place in November 2017 in the Laboratory for Experimental Economics at the University of Bonn. We conducted 8 two-part sessions with 19 to 25 participants each. In sum, we collected data from 171 male participants, mostly students from the university.[13] The first and the second part of the experiment lasted around 45 minutes and 90 minutes, respectively. Participants earned 30 euros on average.

In the first part of the experiment, subjects completed an IQ test and filled out a questionnaire. The second part of the experiment took place one week later, after all subjects had completed the first part, and included the learning exercise as well as the elicitation of

[13] We invited only male subjects as our main research question concerns the consequences of overconfidence, and men are known to be more overconfident than women (Niederle and Vesterlund, 2007). We are not the only study that uses a group of male subjects to investigate overconfidence: see, for example, Burks et al. (2013).

both prior and posterior beliefs.[14] Both parts of the experiment were programmed using zTree (Fischbacher, 2007) and completed by subjects on computers in private cubicles. We describe each part in detail below.

1.2.1. IQ Test and Belief Elicitation. In the first part of the experiment, we evaluated subjects' relative performance in the IQ test, which consisted of 29 standard logic questions. Participants were asked to solve as many of them as possible in 10 minutes. The individual score was calculated based on the number of correctly answered questions minus the number of incorrect answers. To incentivize effort during the test, participants were told that the individual result is important for the next part of the experiment, and their earnings will depend on their scores. After the IQ test, subjects were asked to fill out a questionnaire designed to assess their character traits and individual anxiety levels. At the end of the first session, we reminded participants about the second session one week later, and that they will not be paid unless they show up for the second session.

Between the sessions, we ranked participants according to their IQ test results. For every subject, we calculated his position in the group. The individual position was defined as a number equal to the percent of participants whose test scores were lower or equal to the score obtained by the subject. We defined 20 equilength "performance intervals" ranging from 0% to 100% in steps of 5%. Every participant was assigned the performance interval that his position fell under (with $0-5\%$ denoting the lowest and $95-100\%$ the highest performance interval). We refer to the midpoint of that performance interval as the agent's *relative performance* (denoted by A).

At the beginning of the second session, we elicited subjects' prior belief about their relative position (Confidence I) using the crossover method that is known to be incentive-compatible independently of subjects' risk attitudes (see Schlag et al., 2015). We presented participants with a choice list and asked them to indicate their preferred option in each of the 20 lines. Option A was a lottery with p chance of receiving 5 euros and $1-p$ chance of receiving 0; the winning probability was increasing from $p = 0.05$ to $p = 1$ in 5% steps. Option B stood for a competition with a randomly selected individual, which granted 5 euros if one's IQ test score was higher than their partner, and nothing otherwise. A rational individual would choose

[14] To match subjects' data between the sessions without violating anonymity, we followed a special procedure, which included generating private codes that were used to match subjects to cubicles at the beginning of the second session.

Option A if and only if p is larger than his perceived relative performance. Therefore, we interpret the switching probability as a measure of confidence in one's skills. The procedure was explained to subjects in a simple language, with two examples on how to translate one's beliefs into choices. We followed the same procedure in the second belief elicitation (Confidence II).

1.2.2. Learning Task. After the first belief elicitation, participants completed 6 rounds of a learning task. For every participant, we drew one number for each round, with replacement, from the set $\{-10,-9,...,9,10\}$.[15] We refer to this collection of 6 numbers as an "individual set" and to the set containing all feasible numbers "the feasible set". Participants were reassured that the numbers had been drawn before the experiment started.[16]

In each round, participants were guessing one number taken from their individual set without replacement.[17] For each number, they had to make 4 guesses and enter them into the interface one at a time. After each guess, the computer program calculated a payoff according to the formula:

$$\Pi(e,A,\Phi) = 20 + 0.8 \times (28.6 \times A + \Phi - 0.48\,|e-\Phi|), \tag{1.5}$$

where A denotes the agent's relative performance, Φ is the number drawn, and e refers to his guess. The formula corresponds to the specification of the absolute value loss function. We decided to use this specification because of its simple form and straightforward interpretation. The parameters were chosen such that misguided learning could arise for moderately biased agents. The formula was presented to participants in a descriptive form with an intuitive explanation of the absolute value in terms of distance on the linear scale. We drew

[15] The numbers were drawn from a distribution that put slightly higher weight on numbers in the interval $[-4,4]$. Participants were not presented the exact distribution but were told that the sum of numbers drawn is equal to zero in every round. We explained that some participants had been assigned the number 0, and among the rest half of the participants had been assigned a positive number, while the other half had the same number with the opposite sign.

[16] We informed subjects that the numbers from their individual set had been printed and placed in a sealed envelope next to their seat. They were told not to open the envelopes until the end of the study. As an additional precautionary measure, we placed the envelopes within the sight of the person supervising the session.

[17] We framed the task as "guess the number" instead of "guess your ability and the number", as we aim to test the theory that describes this particular type of situation. We argue that this framing is more suitable to study the implications of overconfidence. In many real-world situations learning about ability is not explicit. For example, when an investor is trading stocks his main task is to generate profits and learn about the market, and not about his ability (even though the profits depend on his analytical skills).

subjects' attention to the fact that the payoff is the higher the closer their guess is to the number drawn (with the highest payoff for the exact match). Participants were informed that, at the end of the experiment, two out of $4 \times 6 = 24$ guesses will be randomly drawn and paid out (with the exchange rate of 0.3).[18]

After entering a guess e, every participant received private feedback. The feedback was equal to one's payoff with an added random component and was displayed on the individual computer screen. The noise was introduced only to ensure that subjects would not be able to infer their ability by matching the feedback to a single identical number in the table.[19]

Participants were informed that they can infer the actual number they are guessing in a given round from their feedback. Knowing their relative performance A, the last guess e, and the payoff Π, they can calculate the unknown number Φ. However, it requires some arithmetical skills. Considering that computational mistakes could influence the learning behavior and obscure the results, we provided subjects with a tool to help them with the task.

1.2.2.1. *Introducing Tables.* Before the learning exercise, every participant was given a set of 21 tables (see Appendix A), from which they could obtain the value of Φ using the feedback they received. The tables contained payoffs for every possible combination of e, Φ, and A. The three parameters jointly determine the payoff, and hence the set of two-dimensional tables contains all feasible payoffs. There is one table for each possible guess e (indicated in the title), the rows indicate the relative performance A (performance intervals are listed in the first column), whereas the columns indicate the number Φ (its values are listed in the first row).

We provided participants with detailed instructions on how to correctly read the tables. Firstly, we described how to find the payoff given e, Φ, and A. A user has to look for a table with his guess in the title, and then look for the intersecting cell corresponding to the row with his relative performance and the column with the number. Secondly, we explained that

[18] One may raise a question whether paying subjects for two elicitation procedures and the learning exercise could induce participants to misreport their beliefs. We admit this possibility, however, we argue that this comment applies only to the second belief elicitation (Confidence II) and does not undermine our main result. Firstly, the instructions for each part were distributed separately, and beforehand participants were not given any information about the remaining tasks. Secondly, in the learning exercise, subjects were not able to influence their payoffs by misreporting their beliefs about themselves. Participants were informed that it is their actual relative performance that enters the payoff function, not their subjective belief.

[19] The random component was drawn from the uniform distribution over the interval $[-0.18, 0.18]$ known to the subjects. Importantly, the noise was not big enough to influence the update. Thus, the set-up can be still described using a model without a random component introduced in Section 1.1.

if someone knows the payoff Π, his last guess e, and his relative performance A, he can obtain the value of Φ by reversing the last steps. After finding the right table, the subject should look at the row with his relative performance and search for his payoff in this row. The column in which lies the payoff indicates the number.

We presented participants with multiple examples and strongly encouraged them to raise questions when in doubt. Every participant had to answer control questions that not only tested their understanding but also pointed out important aspects of the task. Feedback was only displayed after the first guess and participants were not given any information prior to it. Therefore, the first guess that maximizes the expected payoff was $e = 0$. To avoid misunderstandings, we directly told subjects that it is in their best interest to choose zero as their first guess.

1.2.2.2. *Experimental Conditions and Groups.* We introduced two conditions: treatment (we refer to it as "multiple-feedback rounds") and control ("single-feedback rounds"). The two conditions differed with respect to the information provided to participants after each guess. In the multiple-feedback rounds, participants received feedback calculated according to the formula (1.5) after each guess.

In the single-feedback rounds, subjects received feedback calculated according to (1.5) only after their 1^{st} guess. After the 2^{nd} and the 3^{rd} guess computers displayed feedback calculated using the 1^{st} guess in that round. Subjects were notified that no matter what they enter as their 2^{nd} or 3^{rd} guess, the feedback will not reflect their choices. Nevertheless, they were asked to enter their best guess two more times keeping in mind that every guess is equally important for their earnings. By comparing the 3^{rd} and the 4^{th} guess in the multiple-feedback rounds to the corresponding guesses in the single-feedback rounds, one can isolate the effect of informative feedback on misguided learning and prove that the mechanism described in Heidhues et al. (2018) drives the results.

Every participant completed a total of 6 rounds, alternating between the treatment and control conditions. We randomly assigned subjects to two groups (see Table 1.1), with the first group starting with a single-feedback round and the second group starting with a multiple-feedback round.

TABLE 1.1. Experimental conditions and groups.

Round	Group 1	Group 2
1.	SF	MF
2.	MF	SF
3.	SF	MF
4.	MF	SF
5.	SF	MF
6.	MF	SF

SF – single-feedback round
MF – multiple-feedback round

1.3. Results

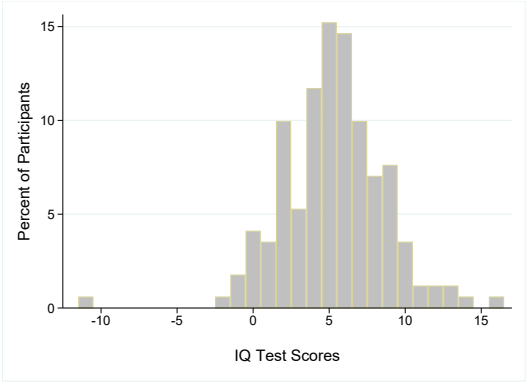
In this section, we present the results of our empirical analysis. Firstly, describe the data on test performance and beliefs, as well as our independent measure of overconfidence. In Section 1.3.2, we look at learning in the multiple-feedback rounds and test the model's predictions for overconfident, underconfident, and unbiased subjects. Next, we use the data from the control condition to exclude alternative explanations of the results. Finally, we discuss learning about ability during the experiment in Section 1.3.3.

1.3.1. IQ Test Results and Elicited Beliefs. In Figure 1.1.(A), we present a histogram of the IQ test results. The scores range from -11 to 16, with over 90% of participants obtaining between 0 and 10 points. The score distribution is fairly symmetrical, with a mean score of 5.29, and a standard deviation of 3.38. In Figure 1.1.(B), we present the distribution of subjects' beliefs about their relative performance elicited before the main task (Confidence I). The average prior belief equals 59.46% and is significantly higher than the average actual position, 55.25% (p-value = 0.092).[20] The average participant is overconfident, yet the magnitude of bias in our sample is not very high. The correlation between subjects' prior beliefs and their actual performance is 0.31 and significant at the 1%-level.

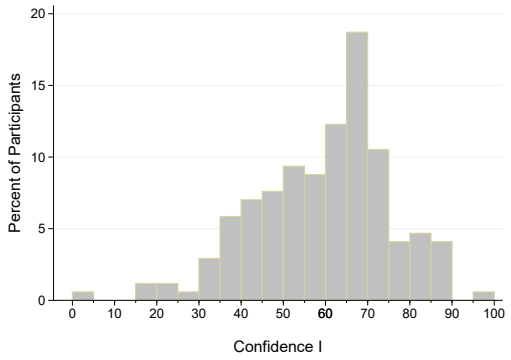
We classify an agent as *overconfident* (*underconfident*) if he assessed his performance to be higher (lower) than his actual position within the group. An unbiased participant correctly estimated his relative performance. As revealed in Confidence I, there are 79 overconfident, 79 underconfident, and 13 unbiased subjects in our sample. On average, underconfident

[20]The average actual position is different from 50% as participants with the same IQ test score were, based on our definition, falling together into one performance interval. We decided not to randomly break ties to avoid misattribution of the result to the random component.

FIGURE 1.1. IQ test results and subjects' beliefs about relative performance.



(A) Distribution of IQ test scores.



(B) Distribution of prior beliefs.

participants held beliefs 20 percentiles lower than their actual position. The average bias of overconfident subjects was around 30 percentiles, meaning that overconfident participants tend to believe that their relative performance was 30 percentiles higher than it actually was. There is a significant difference in the actual performance of overconfident and underconfident subjects. The low-ranked participants tend to overestimate their relative performance, and the high-ranked subjects tend to underestimate it (we address this issue in the following section). The exact values are presented in Table 1.2.

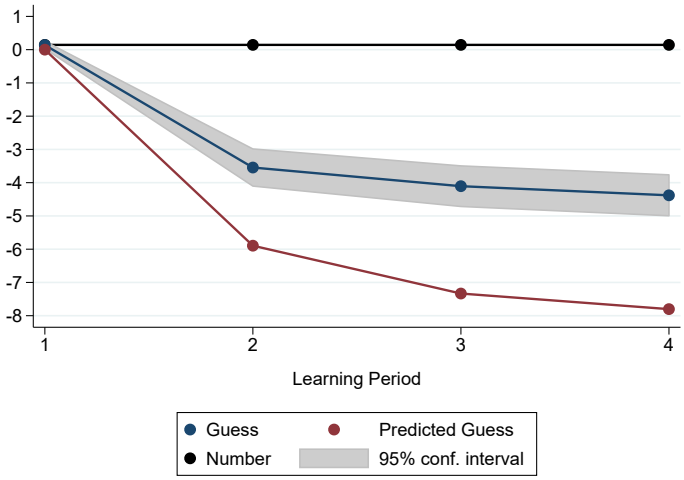
TABLE 1.2. Average performance and beliefs of different types of subjects.

	Underconfident	**Unbiased**	**Overconfident**
Actual Performance:			
Mean (Std. Dev.)	77.50 (16.41)	62.12 (13.14)	31.87 (19.81)
Prior Beliefs:			
Mean (Std. Dev.)	57.31 (16.71)	62.12 (13.14)	61.17 (15.76)

1.3.2. Model Predictions Based on Elicited Beliefs.

1.3.2.1. *Misguided Learning.* First of all, we look at participants' choices in each of the 4 trials. The average guesses of overconfident subjects are presented in Figure 1.2. The blue line connects subjects' actual guesses, and the black points denote the average number being

FIGURE 1.2. Learning of overconfident subjects.

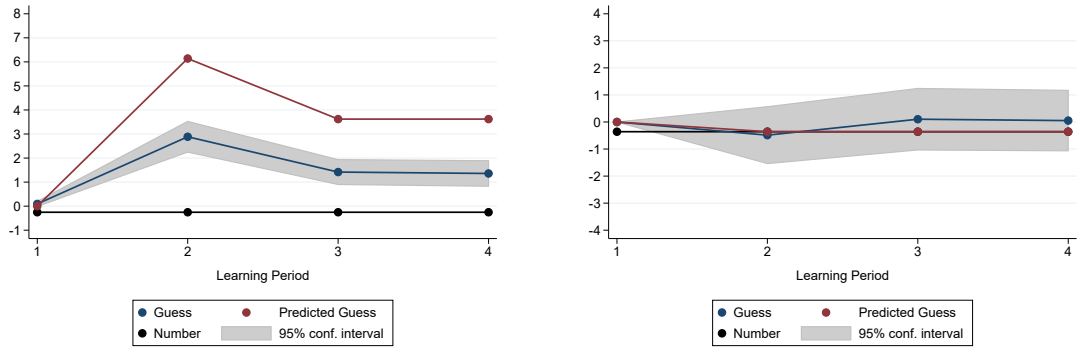


guessed.[21] For every subject, we calculate the 2nd, 3rd and 4th guess predicted by the model, based on the number he was guessing and his bias revealed in Confidence I. The red line connects the average predicted guesses. Although subjects' actual guesses do not coincide with the predicted guesses, the belief path resembles the one predicted by the model. The learning of overconfident agents is self-defeating, with each guess diverging from the true state. We test this formally, by comparing coefficients of a simple regression explaining the difference between a guess and the number with dummy variables, one for each guess (see Table 1.3). The 2nd guess is significantly lower than the 1st guess (one-tailed test: p-value = 0.000), and the 3rd guess is significantly lower than the 2nd guess (one-tailed test: p-value = 0.019).

Although we cannot attest the strict inequality for the 3rd and the 4th guess with similar confidence level, the difference between the 2nd and the 4th guess is highly significant (one-tailed test: p-value = 0.003). Thereby, we confirm the qualitative predictions of the model for overconfident agents. Quantitatively, the effect is around 40% lower than predicted by the model. This may be due to conservatism (under-responsiveness to information known in the literature on asymmetric updating, e.g. Möbius et al., 2022) or subjects learning about their ability over the course of the experiment. We provide evidence for the latter explanation in Section 1.3.3.

[21] Although in every round the sum of numbers given to participants was equal zero, we could not predict the way in which they were distributed among the overconfident, the underconfident and the unbiased agents. Thus, the average of the numbers estimated by different groups was not exactly zero.

FIGURE 1.3. Learning of underconfident and unbiased subjects.



(A) Underconfident subjects.

(B) Unbiased subjects.

The patterns revealed by the underconfident and unbiased agents also follow the model's predictions. For the underconfident agents, the 2nd guess is significantly higher than the 1st guess (one-tailed test: p-value = 0.000), and the 3nd guess is significantly lower than the 2st guess (one-tailed test: p-value = 0.000). After receiving the first feedback, underconfident agents tend to overshoot, but in the following guess they correct their predictions downwards – a pattern also visible in Figure 1.3.(A). Quantitatively, however, the effect is even more mitigated compared to that of the overconfident agents: it is between 53% to 62% lower than predicted by the model. Unbiased agents neither overshoot nor become pessimistic about the state over time, as we can see in Figure 1.3.(B). Their second guess is indistinguishable from the true state, and there is little change in the following trials. Thereby, we confirm Hypothesis 1 for overconfident, underconfident, and unbiased agents (Hypothesis 1.OC, 1.UC, and 1.UB).

1.3.2.2. *Excluding Alternative Explanations.* Before we conclude that participants in our experiment were engaging in misguided learning, we test whether our results were driven by factors outside of the model. For example, the observed patterns might stem from the differences in cognitive ability between underconfident and overconfident subjects, and might not be specific to the environment described in the model. We address this concern with a control condition, in which the main mechanism of the model is switched off.

In the single-feedback rounds, participants received meaningful feedback only after their 1st guess. After the 2nd and the 3rd guess, the number displayed on screen was *independent*

TABLE 1.3. Learning in the multiple-feedback (MF) rounds.

	Overconfident (1)	Unbiased Agents (2)	Underconfident (3)
Dependent variable: difference between a guess and the number in MF rounds. Independent variables: dummy variables for each guess in the MF rounds.			
2^{nd} guess MF	-3.684***	-0.487	2.793***
	(0.342)	(0.570)	(0.381)
3^{rd} guess MF	-4.245***	0.103	1.325***
	(0.391)	(0.466)	(0.291)
4^{th} guess MF	-4.519***	0.051	1.266***
	(0.426)	(0.829)	(0.548)
Const.	-0.004	0.359	0.346
	(0.243)	(0.574)	(0.254)
N	948	156	948

Note: The coefficients at the 2^{nd}, 3^{rd}, and 4^{th} guess MF remain unchanged if we control for subjects' relative performance (their actual position in the IQ test score distribution). Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

of the preceding guess. We kept all other features of the experiment unchanged: as in the multiple-feedback rounds, participants were asked to make four guesses and after each one, a number was displayed on their computer screens. Subjects were informed that the number displayed after the 2^{nd} and the 3^{rd} guess will be based on their 1^{st} guess. Thus, the feedback after the 2^{nd} and the 3^{rd} guess does not bring any new information. The essential feature of the model – the interdependence between actions, feedback, and beliefs – is no longer present, so misguided learning should not arise. However, if there is a downward trend in beliefs of overconfident agents that is independent of the model mechanism, it should be present in the control condition as well.

Firstly, we show that there is no evidence of self-defeating learning in the single-feedback rounds after the second guess. Again, we compare the coefficients of subsequent guesses in a simple regression (Table 1.4). The 3^{rd} guess is not significantly lower than the 2^{nd} guess (one-tailed test: p-value = 0.953), and the 4^{th} guess is not significantly lower than the 3^{rd}

TABLE 1.4. Learning in the single-feedback (SF) rounds.

	Overconfident (1)	Unbiased Agents (2)	Underconfident (3)
Dependent variable: difference between a guess and the number in the SF rounds. Independent variables: dummy variables for each guess in the SF rounds.			
2^{nd} guess SF	-3.350***	0.333	3.493***
	(0.360)	(0.786)	(0.392)
3^{rd} guess SF	-2.958***	0.718	3.080***
	(0.378)	(0.779)	(0.381)
4^{th} guess SF	-2.992***	1.051	3.198***
	(0.361)	(0.828)	(0.387)
Const.	0.278	-0.513	-0.118
	(0.258)	(0.749)	(0.237)
N	948	156	948

Note: The coefficients at the 2^{nd}, 3^{rd}, and 4^{th} guess SF remain unchanged if we control for subjects' relative performance (their actual position in the IQ test score distribution). Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

guess (one-tailed test: p-value = 0.431). The results prove that, for overconfident agents, the belief path in the single-feedback rounds does not exhibit the pattern characteristic of self-defeating learning. While we cannot reject the hypothesis that the learning process of underconfident agents is self-correcting in the single-feedback rounds, the extent of correction is much lower. In the multiple-feedback rounds, participants corrected 55% of the initial overshooting, and the correction in the single-feedback rounds did not exceed 7%.

Secondly, we pool the data from the single- and multiple-feedback rounds and look at the effect of receiving informative feedback on learning. We regress the difference between a subject's guess and the number on a dummy variable indicating a multiple-feedback round. The results are gathered in Table B.2 in Appendix B. For overconfident participants, being in a multiple-feedback round increases the negative difference between a guess and the number by -1.57 in the 3^{rd} guess (one-tailed test: p-value = 0.000) and by -1.81 in the 4^{th} guess

(one-tailed test: p-value = 0.000).[22] Informative feedback makes overconfident subjects more mistaken both in the 3rd and in the 4th guess.[23] As a final test, we regress the difference between the 4th and the 2nd guess on a dummy variable indicating a multiple-feedback round (see Table B.4 in Appendix B). The coefficient is negative and highly significant: providing overconfident subjects with informative feedback shifts their beliefs downwards by −1.19, which constitutes 67% of the effect predicted by the model. Moreover, receiving informative feedback affects underconfident but not unbiased subjects, in line with the model predictions. The results for underconfident and unbiased agents are delegated to Appendix B.

1.3.2.3. *Individual Heterogeneity.* In this section, we analyze how misguided learning depends on subjects' bias. To this end, we conduct a regression analysis similar to the one presented in the previous section, but we allow for the effect to depend on the bias (the degree of over- and under-confidence measured before the task). The results are gathered in Table 1.5. The dependent variable is the difference between the subject's guess and the number in the multiple-feedback rounds, whereas independent variables include dummy variables indicating consecutive guesses and their interactions with a measure of agents' bias. "Bias" variable takes values between −1 and 1, with positive (negative) values characterizing overconfident (underconfident) subjects. We analyze separately the behavior of overconfident and underconfident subjects. However, this time, we include unbiased agents in each group, as they provide a useful benchmark for studying the effect of bias (similar regressions without unbiased subjects could be found in Table B.5 in Appendix B).

The coefficients at the interaction terms provide evidence for a significant effect of bias on the learning process. For the overconfident subjects, a 10-percentile increase in bias exacerbates mislearning by −0.68, −0.74, and −0.76 in the 2nd, 3rd, and 4th guess, respectively. Thus, we confirm Hypothesis 2.OC that more overconfident participants tend to form more pessimistic beliefs and end up further away from the true state compared to less overconfident subjects. Moreover, we observe a similar effect in the group of underconfident subjects. A 10-percentile increase in bias results in additional overestimation of the number by 0.65 in the 2nd guess (underconfident agents' bias takes negative values, hence the effect goes in the

[22] While the differences might appear small, they are close to the values predicted by the model (−1.59 in the 3rd guess and −2.05 in the 4th guess).

[23] The negative sign indicates that overconfident subjects became more pessimistic about the state.

TABLE 1.5. The effect of bias on learning in the multiple-feedback rounds.

	Overconfident or Unbiased (1)		Underconfident or Unbiased (2)	
Dependent variable: the difference between a guess and the number in MF rounds. Independent variables: dummy variables for each guess and their interactions.				
2^{nd} guess MF	-1.509***	(0.438)	1.198**	(0.521)
3^{rd} guess MF	-1.758***	(0.467)	0.225	(0.383)
4^{th} guess MF	-1.961***	(0.524)	0.179	(0.419)
Bias	-0.914	(1.113)	-1.212	(1.461)
Bias × 2^{nd} guess MF	-6.848***	(2.089)	-6.526**	(2.642)
Bias × 3^{rd} guess MF	-7.440***	(2.001)	-5.346***	(1.783)
Bias × 4^{th} guess MF	-7.600***	(2.227)	-5.278***	(1.914)
Const.	0.277	(0.350)	0.138	(0.372)
N	1104		1104	

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

predicted direction). It confirms Hypothesis 2.UC, as more underconfident participants end up further away from the true state after the first feedback.

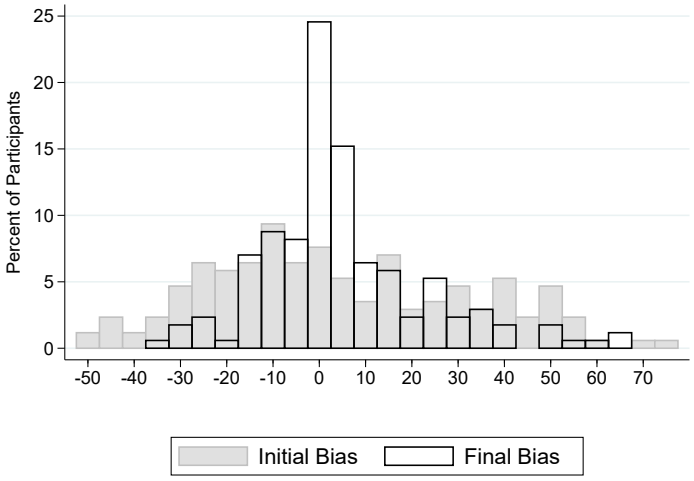
The results also shed light on Hypothesis 3.UC&OC. While we admit that our setting is more suitable to test the short-term dynamics of the model, we argue that the last guess is informative about the long-term. In our setting, most participants are expected to reach the stable belief within 4 trials.[24] The average stable belief is 3.78 for underconfident and −8.42 for overconfident subjects – both values are very close to the average predicted 4^{th} guess of 3.62 and −7.80, respectively. For this reason, we treat the 4^{th} guess as close enough to the stable belief and test Hypothesis 3.UC&OC. The coefficient at the interaction with the 4^{th} guess in Table 1.5 informs us about the effect of bias on the end belief. Both for the underconfident

[24]It is due to the chosen parameters, as well as the discrete action space (if we did not require subjects' guesses to be integers, the convergence to the stable belief would take longer than 4 periods).

and overconfident subjects, more biased individuals end up further away from the true state, in line with the model predictions.

1.3.3. Learning about Ability. As we have already mentioned, there is a substantial gap between subjects' guesses and the decisions predicted by the model. Why is misguided learning less pronounced than the model predicts? While the model is based on the assumption that agents do not change their beliefs about ability, we did not impose this assumption on our subjects.[25] As a result, we observe learning about ability over the course of the experiment. In Figure 1.4, we present the distributions of participants' bias before and after the task (based on Confidence I and Confidence II).

FIGURE 1.4. Distribution of subjects' bias before and after the task.



The average bias of underconfident subjects decreased from 20.2 to 6 percentiles after the learning exercise, and the average bias of the overconfident subject decreased from 29.4 to 16.5 percentiles. The changes in mean beliefs are statistically significant both for the overconfident and the underconfident subjects. After the learning exercise, 18% of overconfident and 30% of underconfident subjects became unbiased. Nevertheless, Confidence II reveals

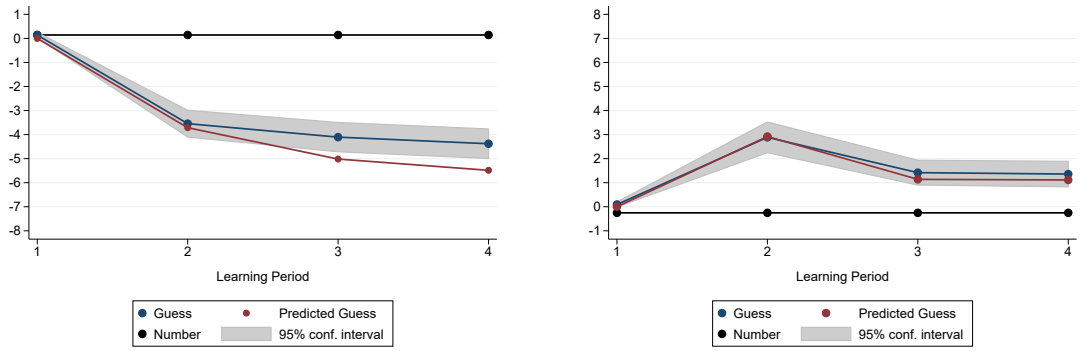
[25] It was our intention from the beginning to leave participants with an opportunity to revise their beliefs about their cognitive ability. We believe that imposing too many restrictions on subjects' behavior would make the test meaningless, as it would tell us little about how subjects would behave if not restricted. In our view, this design provides a more powerful and interesting test of the theory. Our results show that even without requiring subjects to hold on to their initial assessment, they follow the theoretical predictions as they *choose* to stick to their biased beliefs about their cognitive ability.

that a significant portion of the sample held incorrect beliefs even after the learning exercise, and many of them were engaging in misguided learning till the very last round.

The data from Confidence I and II tells us little about the changes in subjects' beliefs about ability *during* the learning exercise. Fortunately, the experimental design enables us to divulge the beliefs about one's relative performance with few additional assumptions (see Appendix C.1). We assume that participants update their beliefs about ability at the beginning of each round, and use their initial guesses to obtain a measure of those updated beliefs (we use the 2^{nd} guess, as in the 1^{st} guess subjects were instructed to enter 0). The round-to-round changes in subjects' beliefs about ability are described in detail in Appendix C.2. Here, let us only point out their implications for the model's performance. We use beliefs revealed from the 2^{nd} guess to calculate the model's predictions for the 3^{rd} and the 4^{th} guess. The results for the underconfident and overconfident agents are presented in Figure 1.5.

The average predicted 3^{rd} and 4^{th} guess (the red line) is now much closer to the average actual guess (the blue line). The better fit is reflected in the estimates of how well the model fits the data. The model based on revealed beliefs explains 73.5% variation in the choice data (the 3^{rd} and the 4^{th} guess), compared to 52.3% if we use its predictions based on elicited beliefs (see Appendix C.3). We conclude that the difference between our initial theoretical predictions and the actual guesses is due to participants learning their ability during the task. If we control for changes in beliefs from round to round, the model closely tracts subjects' behavior.

FIGURE 1.5. Model's predictions based on revealed beliefs (MF rounds).



(A) Overconfident subjects.

(B) Underconfident subjects.

1.4. Comparison with Ego-neutral Environment

We hypothesize that our results are driven in part by participants' tendency to interpret feedback in a self-serving manner. We designed an additional control condition to test whether motivated reasoning is driving our results. In this condition, participants were learning about two parameters that were *both* ego-neutral. We used the same experimental design, with the only difference being that subjects performed the main task based on the performance parameter of another subject.[26] We assume that the performance of another individual is irrelevant to one's ego. Participants were informed that each of them will be randomly matched to another subject who completed the same IQ test and revealed similar beliefs through the same elicitation procedure. Before the main task, we elicited subjects' beliefs about the relative performance of the subject matched to them and distinguished overconfident, underconfident, and unbiased agents (with respect to their partner's performance). We again elicited beliefs about the performance of the matched partner after the task.

We collected data from 151 male participants, mostly students from the University of Bonn. There is no significant difference between the two groups in relative performance nor initial bias about own performance (see Table 1.6). In the control group, 73 subjects were classified as overconfident about their partners' performance, 73 as underconfident, and 9 as unbiased. In what follows, we refer to a control subject as overconfident (underconfident) if he overestimated (underestimated) his match's performance.[27]

TABLE 1.6. Differences between participants in the two conditions.

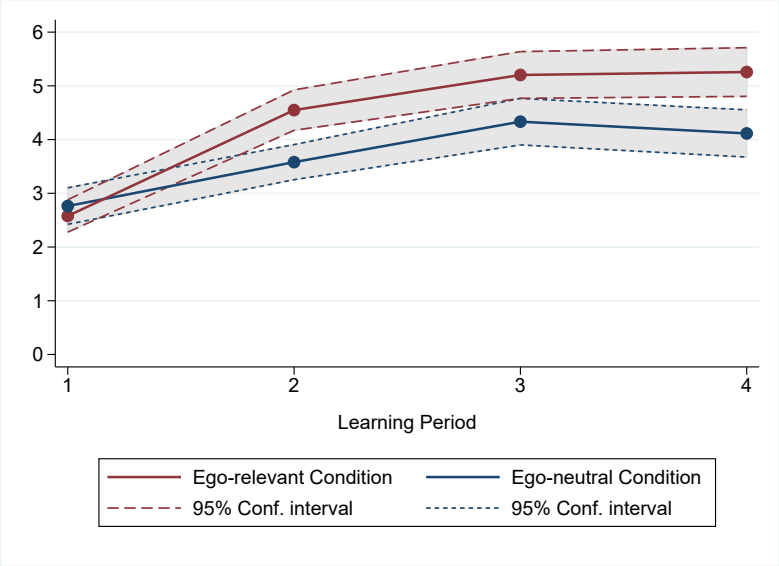
	Ego-neutral	**Ego-relevant**		Diff < 0	Diff ≠ 0	Diff > 0
Performance	0.579	0.552	p-value:	0.780	0.401	0.200
	(0.023)	(0.022)				
Initial Bias	0.014	0.042	p-value:	0.180	0.360	0.820
	(0.022)	(0.021)				
N	155	171				

[26] The almost identical experimental design enables to control for possible confounds such as, for example, the way subjects' attention was directed during the experiment.

[27] One consequence of the random assignment of partners in the ego-neutral condition is that the average performance of overconfident subjects in the ego-neutral condition (overconfidence defined with respect to the other's performance) is higher than that of the overconfident subjects in the ego-relevant condition (overconfidence defined with respect to own performance). See Appendix D.2 for details. We address this problem by controlling for the performance of the decision-maker and his initial bias.

1.4.1. Learning about the Number. Figure 1.6 presents the average distance between the overconfident agent's guess and the estimated number in the two conditions (as a measure of distance we use the absolute difference between a guess and the number). The distance is larger in the ego-relevant condition, that is, for agents whose feedback was based on their own relative performance. Table 1.7 presents the results of a corresponding regression analysis. The distance between the agent's last guess and the number is larger by 1.14 in the ego-relevant condition. The effect persists when we control for the relative performance of the decision-maker (the second column in Table 1.7) or his initial bias and relative performance (the third column in Table 1.7).[28] Overconfident participants in the ego-relevant condition end up *more* mistaken about the state of the world compared to similar subjects in the ego-neutral condition. In the last two columns in Table 1.7, we test for the treatment effect using the nearest neighbor matching estimator. In Specification 4, we match participants based on the relative performance of the decision-maker, and in Specification 5, based on the initial bias and relative performance. Both specifications yield similar results.[29]

FIGURE 1.6. Distance between a guess and the number in the ego-relevant and the ego-neutral condition.



[28] Similar regressions for the 2nd and the 3rd guess are presented in Appendix D.3.

[29] As a final test, we add to specifications 1-3 a control for the model's predictions (decisions implied by the model). The coefficients at the "Ego-relevant" variable remain similar and highly significant. We report them in Appendix D.3.

TABLE 1.7. The effect of ego-relevance on learning of overconfident agents.

	Dependent variable: the absolute difference between the 4^{th} guess and the number.				
	(1)	(2)	(3)	(4)	(5)
Ego-relevant	1.143^{**}	1.698^{***}	1.630^{***}	1.570^{***}	1.229^{***}
	(0.539)	(0 .522)	(0.492)	(0.382)	(0.375)
Controls 1	No	Yes	Yes		
Controls 2	No	No	Yes		
Adjustment Type	Regression	Regression	Regression	Matching	Matching
Observations	456	456	456	456	456

Note: The dependent variable is the absolute difference between the 4^{th} guess and the number. The sample includes only overconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. In the matching estimator, observations are matched to the nearest neighbor based on the relative performance (Specification 4), and the initial bias and relative performance (Specification 5). In Specification 1-3, standard errors clustered at the individual level. In Specification 4-5, consistent standard errors as in Abadie and Imbens (2006). Their values in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

What makes participants in the ego-relevant condition more mistaken about the state of the world? In the ego-relevant condition, overconfident agents might be reluctant to abandon their model of the world, as it would require them to admit to lower performance. Consequently, they will be less willing to correct their guesses compared to overconfident agents in the ego-neutral condition.[30]

Our hypothesis finds support in the data from underconfident agents. In the ego-relevant condition, underconfident participants tend to overshoot significantly less compared to similarly underconfident subjects in the ego-neutral condition. The effect is highly significant even after controlling for the initial bias and relative performance (see Table 1.8). The sign of the effect is opposite to that of the overconfident agents – the ego-relevance of the task makes underconfident agents *less* misguided. However, the direction is consistent with motivated reasoning: in the ego-relevant condition, underconfident agents are more willing to abandon

[30]Nonetheless, misguided learning is not entirely eliminated in the control condition, pointing towards the role of biased beliefs as its main source (the analysis of the control data analogous to Section 1.3.2 could be found in Appendix D.1). Our results suggest that misguided learning can emerge in ego-neutral settings, although it is not as pronounced if agents are more willing to update their beliefs.

TABLE 1.8. The effect of ego-relevance on learning of underconfident agents.

	Dependent variable: the absolute difference between the 2^{nd} guess and the number.				
	(1)	(2)	(3)	(4)	(5)
Ego-relevant	-0.849**	-1.099***	-0.976***	-1.056***	-0.816***
	(0.403)	(0.393)	(0.367)	(0.314)	(0.312)
Controls 1	No	Yes	Yes		
Controls 2	No	No	Yes		
Adjustment Type	Regression	Regression	Regression	Matching	Matching
Observations	456	456	456	456	456

Note: The dependent variable is the absolute difference between the 2^{nd} guess and the number. The sample includes only underconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. In the matching estimator, observations are matched to the nearest neighbor based on the relative performance (Specification 4), and the initial bias and relative performance (Specification 5). In Specification 1-3, standard errors clustered at the individual level. In Specification 4-5, consistent standard errors as in Abadie and Imbens (2006). Their values in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

their previously held beliefs, as it allows them to admit that they performed better than expected. This interpretation is also supported by the data on learning about ability presented in the next section.

1.4.2. Learning about Own versus Other's Ability. The data from the second belief elicitation (Confidence II) reveals that, in the ego-neutral condition, 33 participants became unbiased about the ability of their match (compared to 38 participants in the ego-relevant condition). While the fraction of subjects who became unbiased is almost the same in the two conditions, the composition of types differs. In the ego-relevant condition, 30% of underconfident and 18% of overconfident participants revealed unbiased beliefs about their ability after the task. In the ego-neutral condition, these proportions are reversed: 18% of underconfident and 27% of overconfident subjects were classified as unbiased after the task.

The results in Table 1.9 demonstrate that overconfident participants in the ego-relevant condition are less likely to become unbiased compared to similarly overconfident participants in the ego-neutral control. At the same time, underconfident participants are more likely to become unbiased when learning about their own ability. Importantly, the effect is

TABLE 1.9. The effect of ego-relevance on becoming unbiased after the task.

	Dependent variable: binary variable indicating whether subject became unbiased after the task.					
	Overconfident			*Underconfident*		
	(1)	(2)	(3)	(1)	(2)	(3)
Ego-relevant	-0.097	-0.168**	-0.147*	0.126*	0.159**	0.165**
	(0.068)	(0.070)	(0.075)	(0.069)	(0.072)	(0.075)
Controls 1	No	Yes	Yes	No	Yes	Yes
Controls 2	No	No	Yes	No	No	Yes
Observations	152	152	152	152	152	152

Note: The dependent variable is a binary variable indicating whether subject became unbiased after the task, as revealed in Confidence II. "Ego-relevant" indicates assignment to the ego-relevant condition. Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker.

Standard errors clustered at the individual level. Their values in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

present if we control for the relative performance of the decision-maker (the second specification) or his initial bias and relative performance (the third specification). The sign of the effect is indicative of motivated reasoning: overconfident subjects are less inclined to learn that they performed worse than expected, and underconfident subjects are more inclined to learn that they did better.

1.4.3. Discussion: Learning about Multiple Parameters. Our results show that self-defeating learning is more likely to arise and persist when one's ego is at stake. Overconfident participants, reluctant to revise their beliefs about ability downwards, are bound to become mistaken about the state of the world. On the other hand, underconfident agents are *more* willing to correct their beliefs about their own ability, making them *less* susceptible for mislearning in ego-relevant settings. The results also indicate that, when learning involves multiple parameters and some of them are ego-relevant, people will be steered to learn along the dimension that brings them higher ego utility. In the case of overconfident agents, this means holding onto their inflated beliefs, while for underconfident agents – revising them upwards. Still, as we have seen, neither underconfident nor overconfident subjects go the full length in updating or holding on to their biased beliefs about ability. More research is needed to understand

how people make the trade-off between ego utility and the expected benefit of learning the state.

1.5. Conclusions

Successful decision-making often requires forming beliefs about various characteristics of the environment. However, learning about multiple parameters is rarely independent: the way an agent updates his beliefs about one aspect might influence his reasoning about other parameters. In particular, if the agent overestimates his ability, he may repeatedly misinterpret the data and fail to take the optimal action time after time, falling into a vicious circle of misguided learning. In this paper, we experimentally test subjects' propensity to engage in this kind of behavior. The results corroborate the theory formulated by Heidhues et al. (2018) and demonstrate that misguided learning is a real-world phenomenon that is likely to afflict biased agents. As long as people hold on to their overconfident beliefs, they will continue to misread the data and form erroneous beliefs about their environment. The problem is aggravated when agents hold overconfident beliefs about characteristics they care about: their reluctance to revise their beliefs downwards exacerbates the tendency to mislearn. Allowing agents to experiment and acquire new information is, in these cases, counterproductive.

APPENDIX A

The Use of Tables by Biased Agents

In the following section we show how a myopic agent who only updates his beliefs about the state of the world uses the tables in the multiple- and single-feedback rounds.

In the first example, we assume that the agent's relative performance parameter is $A = 47.5\%$ and he is guessing the number $\Phi = -1$ in a multiple-feedback round. The agent is *overconfident* and believes that his performance lies in the $55-60\%$ interval. Figure A.1 illustrates this case: we depicted the agent's actual performance and the number in red, and the agent's beliefs and actions in blue. The agent enters $e_1 = 0$ as his first guess. Afterwards, the computer displays the feedback of 29.71, which consists of the payoff $\Pi_1 = 29.68$ and the added random component $\epsilon_1 = 0.03$. The agent believes that his relative performance lies in the $55-60\%$ interval, therefore he looks at the row outlined in blue, and searches for a value that is the closest to his feedback. There is only one such value (29.60), and the agent concludes that the number he is guessing is equal to $\phi_2 = -3$. The agent updates his beliefs about the number and enters $e_2 = -3$ as his second guess. The computer displays a new feedback: 29.45. The agent browses the tables looking for the one with the number -3 in the title (see Figure A.2). Once again he looks at the row with the relative performance between 55% and 60% and compares his feedback to the values in that row. The overconfident agent concludes that the number must be equal to $\phi_3 = -4$ and he chooses $e_3 = -4$ as his third guess. In the following step, he becomes even more mistaken, concluding that the number is $\phi_4 = -5$ and choosing $e_4 = -5$ as his last guess (presented in Figure A.3). The overconfident agent's beliefs change in the following way: $\phi_1 = 0$, $\phi_2 = -3$, $\phi_3 = -4$, $\phi_4 = -5$. As predicted by the model, the learning process is self-defeating: the additional feedback drives the agent's beliefs further away from the true state.

In a single-feedback round, the agent's reasoning after the first guess is the same as in the multiple-feedback round. He forms a belief $\phi_2 = -3$ and enters the optimal action $e_2 = -3$. In contrast to the multiple-feedback round, any feedback the agent receives afterward is based on his first guess, hence he should use the table with 0 in the title. The agent receives the

feedback 29.59 (the noise component is $\epsilon_2 = -0.09$). The closest value in the table is again 29.68, so he should enter $e_3 = -3$. The last feedback differs only with respect to the noise term, inducing a belief $\phi_4 = -3$ and prompting the action $e_4 = -3$. In the single-feedback rounds, the agent's beliefs change as follows: $\phi_1 = 0$, $\phi_2 = -3$, $\phi_3 = -3$, $\phi_4 = -3$. Severing the link between the actions and output precludes self-defeating learning.

The next example considers an *underconfident* agent with the relative performance $A = 62.5\%$ who is guessing the number $\Phi = 4$ in a multiple-feedback round. The agent believes that his relative performance is 10% lower and lies in the $50-55\%$ interval. When he sees the feedback of 35.85 (the actual payoff 35.96 with the added noise term $\epsilon_1 = -0.11$), he infers that the number is equal to $\phi_2 = 9$. We depict the first step in Figure A.4. The agent's actual performance parameter and the number are in red, and his beliefs and choices are in blue. The underconfident agent enters $e_2 = 9$ as his second guess and obtains the feedback 35.57 that includes the noise term $\epsilon_2 = -0.01$. He goes to the table with the number 9 in the title (presented in Figure A.5). The value closest to his feedback, i.e. $\Pi = 35.66$, points to the number $\phi_3 = 6$. The agent updates his beliefs, enters the optimal action $e_3 = 6$ and receives the feedback of 36.78 ($\epsilon_3 = 0.05$). In the last step, he turns to table 6 (presented in Figure A.6), from which he infers that $\phi_4 = 6$ is the number he is looking for, thus he enters $e_4 = 6$. The underconfident agent's beliefs follow the path: $\phi_1 = 0$, $\phi_2 = 9$, $\phi_3 = 6$, $\phi_4 = 6$. As predicted by the model, the underconfident agent first overshoots and then corrects his actions. In a single-feedback round, the agent would not update his beliefs after the second guess, thus entering $e_3 = e_4 = 9$ as his third and fourth guess.

The last example illustrates the behavior of an *unbiased* agent, who has the relative performance of $A = 72.5\%$ and is guessing the number $\Phi = -4$ in a multiple-feedback round. After entering $e_1 = 0$ the agent receives the feedback of 31.82 (the actual payoff is 31.85 and the added noise term $\epsilon_1 = -0.03$), which points to the correct number $\phi_2 = -4$. The agent enters $e_2 = -4$ and turns to the table with -4 in the title (presented in Figure A.7). The feedback displayed on his screen is the payoff of 33.39 with a perturbation, which points to the number $\phi_3 = -4$. Regardless of the noise realization, the feedback will not be closer to any other value but 33.39. The agent chooses the optimal action $e_4 = -4$ as his fourth guess. The learning process of the unbiased individual is immediate and his belief is stable afterward.

Ihre Schätzung war: 0

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	30.47	31.65	32.84	34.02	35.20	36.39	37.57	38.76	39.94	41.12	42.31	42.72	43.14	43.56	43.97	44.39	44.80	45.22	45.64	46.05	46.47
90 – 95%	29.32	30.51	31.69	32.88	34.06	35.24	36.43	37.61	38.80	39.98	41.16	41.58	42.00	42.41	42.83	43.24	43.66	44.08	44.49	44.91	45.32
85 – 90%	28.18	29.36	30.55	31.73	32.92	34.10	35.28	36.47	37.65	38.84	40.02	40.44	40.85	41.27	41.68	42.10	42.52	42.93	43.35	43.76	44.18
80 – 85%	27.04	28.22	29.40	30.59	31.77	32.96	34.14	35.32	36.51	37.69	38.88	39.29	39.71	40.12	40.54	40.96	41.37	41.79	42.20	42.62	43.04
75 – 80%	25.89	27.08	28.26	29.44	30.63	31.81	33.00	34.18	35.36	36.55	37.73	38.15	38.56	38.98	39.40	39.81	40.23	40.64	41.06	41.48	41.89
70 – 75%	24.75	25.93	27.12	28.30	29.48	30.67	31.85	33.04	34.22	35.40	36.59	37.00	37.42	37.84	38.25	38.67	39.08	39.50	39.92	40.33	40.75
65 – 70%	23.60	24.79	25.97	27.16	28.34	29.52	30.71	31.89	33.08	34.26	35.44	35.86	36.28	36.69	37.11	37.52	37.94	38.36	38.77	39.19	39.60
60 – 65%	22.46	23.64	24.83	26.01	27.20	28.38	29.56	30.75	31.93	33.12	34.30	34.72	35.13	35.55	35.96	36.38	36.80	37.21	37.63	38.04	38.46
55 – 60%	21.32	22.50	23.68	24.87	26.05	27.24	28.42	29.60	30.79	31.97	33.16	33.57	33.99	34.40	34.82	35.24	35.65	36.07	36.48	36.90	37.32
50 – 55%	20.17	21.36	22.54	23.72	24.91	26.09	27.28	28.46	29.64	30.83	32.01	32.43	32.84	33.26	33.68	34.09	34.51	34.92	35.34	35.76	36.17
45 – 50%	19.03	20.21	21.40	22.58	23.76	24.95	26.13	27.32	28.50	29.68	30.87	31.28	31.70	32.12	32.53	32.95	33.36	33.78	34.20	34.61	35.03
40 – 45%	17.88	19.07	20.25	21.44	22.62	23.80	24.99	26.17	27.36	28.54	29.72	30.14	30.56	30.97	31.39	31.80	32.22	32.64	33.05	33.47	33.88
35 – 40%	16.74	17.92	19.11	20.29	21.48	22.66	23.84	25.03	26.21	27.40	28.58	29.00	29.41	29.83	30.24	30.66	31.08	31.49	31.91	32.32	32.74
30 – 35%	15.60	16.78	17.96	19.15	20.33	21.52	22.70	23.88	25.07	26.25	27.44	27.85	28.27	28.68	29.10	29.52	29.93	30.35	30.76	31.18	31.60
25 – 30%	14.45	15.64	16.82	18.00	19.19	20.37	21.56	22.74	23.92	25.11	26.29	26.71	27.12	27.54	27.96	28.37	28.79	29.20	29.62	30.04	30.45
20 – 25%	13.31	14.49	15.68	16.86	18.04	19.23	20.41	21.60	22.78	23.96	25.15	25.56	25.98	26.40	26.81	27.23	27.64	28.06	28.48	28.89	29.31
15 – 20%	12.16	13.35	14.53	15.72	16.90	18.08	19.27	20.45	21.64	22.82	24.00	24.42	24.84	25.25	25.67	26.08	26.50	26.92	27.33	27.75	28.16
10 – 15%	11.02	12.20	13.39	14.57	15.76	16.94	18.12	19.31	20.49	21.68	22.86	23.28	23.69	24.11	24.52	24.94	25.36	25.77	26.19	26.60	27.02
5 – 10%	9.88	11.06	12.24	13.43	14.61	15.80	16.98	18.16	19.35	20.53	21.72	22.13	22.55	22.96	23.38	23.80	24.21	24.63	25.04	25.46	25.88
0 – 5%	8.73	9.92	11.10	12.28	13.47	14.65	15.84	17.02	18.20	19.39	20.57	20.99	21.40	21.82	22.24	22.65	23.07	23.48	23.90	24.32	24.73

FIGURE A.1. The use of tables by an overconfident agent: the 2nd guess.

Ihre Schätzung war: −3

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	31.62	32.80	33.99	35.17	36.36	37.54	38.72	39.91	40.32	40.74	41.16	41.57	41.99	42.40	42.82	43.24	43.65	44.07	44.48	44.90	45.32
90 – 95%	30.48	31.66	32.84	34.03	35.21	36.40	37.58	38.76	39.18	39.60	40.01	40.43	40.84	41.26	41.68	42.09	42.51	42.92	43.34	43.76	44.17
85 – 90%	29.33	30.52	31.70	32.88	34.07	35.25	36.44	37.62	38.04	38.45	38.87	39.28	39.70	40.12	40.53	40.95	41.36	41.78	42.20	42.61	43.03
80 – 85%	28.19	29.37	30.56	31.74	32.92	34.11	35.29	36.48	36.89	37.31	37.72	38.14	38.56	38.97	39.39	39.80	40.22	40.64	41.05	41.47	41.88
75 – 80%	27.04	28.23	29.41	30.60	31.78	32.96	34.15	35.33	35.75	36.16	36.58	37.00	37.41	37.83	38.24	38.66	39.08	39.49	39.91	40.32	40.74
70 – 75%	25.90	27.08	28.27	29.45	30.64	31.82	33.00	34.19	34.60	35.02	35.44	35.85	36.27	36.68	37.10	37.52	37.93	38.35	38.76	39.18	39.60
65 – 70%	24.76	25.94	27.12	28.31	29.49	30.68	31.86	33.04	33.46	33.88	34.29	34.71	35.12	35.54	35.96	36.37	36.79	37.20	37.62	38.04	38.45
60 – 65%	23.61	24.80	25.98	27.16	28.35	29.53	30.72	31.90	32.32	32.73	33.15	33.56	33.98	34.40	34.81	35.23	35.64	36.06	36.48	36.89	37.31
55 – 60%	22.47	23.65	24.84	26.02	27.20	28.39	29.57	30.76	31.17	31.59	32.00	32.42	32.84	33.25	33.67	34.08	34.50	34.92	35.33	35.75	36.16
50 – 55%	21.32	22.51	23.69	24.88	26.06	27.24	28.43	29.61	30.03	30.44	30.86	31.28	31.69	32.11	32.52	32.94	33.36	33.77	34.19	34.60	35.02
45 – 50%	20.18	21.36	22.55	23.73	24.92	26.10	27.28	28.47	28.88	29.30	29.72	30.13	30.55	30.96	31.38	31.80	32.21	32.63	33.04	33.46	33.88
40 – 45%	19.04	20.22	21.40	22.59	23.77	24.96	26.14	27.32	27.74	28.16	28.57	28.99	29.40	29.82	30.24	30.65	31.07	31.48	31.90	32.32	32.73
35 – 40%	17.89	19.08	20.26	21.44	22.63	23.81	25.00	26.18	26.60	27.01	27.43	27.84	28.26	28.68	29.09	29.51	29.92	30.34	30.76	31.17	31.59
30 – 35%	16.75	17.93	19.12	20.30	21.48	22.67	23.85	25.04	25.45	25.87	26.28	26.70	27.12	27.53	27.95	28.36	28.78	29.20	29.61	30.03	30.44
25 – 30%	15.60	16.79	17.97	19.16	20.34	21.52	22.71	23.89	24.31	24.72	25.14	25.56	25.97	26.39	26.80	27.22	27.64	28.05	28.47	28.88	29.30
20 – 25%	14.46	15.64	16.83	18.01	19.20	20.38	21.56	22.75	23.16	23.58	24.00	24.41	24.83	25.24	25.66	26.08	26.49	26.91	27.32	27.74	28.16
15 – 20%	13.32	14.50	15.68	16.87	18.05	19.24	20.42	21.60	22.02	22.44	22.85	23.27	23.68	24.10	24.52	24.93	25.35	25.76	26.18	26.60	27.01
10 – 15%	12.17	13.36	14.54	15.72	16.91	18.09	19.28	20.46	20.88	21.29	21.71	22.12	22.54	22.96	23.37	23.79	24.20	24.62	25.04	25.45	25.87
5 – 10%	11.03	12.21	13.40	14.58	15.76	16.95	18.13	19.32	19.73	20.15	20.56	20.98	21.40	21.81	22.23	22.64	23.06	23.48	23.89	24.31	24.72
0 – 5%	9.88	11.07	12.25	13.44	14.62	15.80	16.99	18.17	18.59	19.00	19.42	19.84	20.25	20.67	21.08	21.50	21.92	22.33	22.75	23.16	23.58

FIGURE A.2. The use of tables by an overconfident agent: the 3rd guess.

Ihre Schätzung war: −4

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	32.00	33.19	34.37	35.56	36.74	37.92	39.11	39.52	39.94	40.36	40.77	41.19	41.60	42.02	42.44	42.85	43.27	43.68	44.10	44.52	44.93
90 – 95%	30.86	32.04	33.23	34.41	35.60	36.78	37.96	38.38	38.80	39.21	39.63	40.04	40.46	40.88	41.29	41.71	42.12	42.54	42.96	43.37	43.79
85 – 90%	29.72	30.90	32.08	33.27	34.45	35.64	36.82	37.24	37.65	38.07	38.48	38.90	39.32	39.73	40.15	40.56	40.98	41.40	41.81	42.23	42.64
80 – 85%	28.57	29.76	30.94	32.12	33.31	34.49	35.68	36.09	36.51	36.92	37.34	37.76	38.17	38.59	39.00	39.42	39.84	40.25	40.67	41.08	41.50
75 – 80%	27.43	28.61	29.80	30.98	32.16	33.35	34.53	34.95	35.36	35.78	36.20	36.61	37.03	37.44	37.86	38.28	38.69	39.11	39.52	39.94	40.36
70 – 75%	26.28	27.47	28.65	29.84	31.02	32.20	33.39	33.80	34.22	34.64	35.05	35.47	35.88	36.30	36.72	37.13	37.55	37.96	38.38	38.80	39.21
65 – 70%	25.14	26.32	27.51	28.69	29.88	31.06	32.24	32.66	33.08	33.49	33.91	34.32	34.74	35.16	35.57	35.99	36.40	36.82	37.24	37.65	38.07
60 – 65%	24.00	25.18	26.36	27.55	28.73	29.92	31.10	31.52	31.93	32.35	32.76	33.18	33.60	34.01	34.43	34.84	35.26	35.68	36.09	36.51	36.92
55 – 60%	22.85	24.04	25.22	26.40	27.59	28.77	29.96	30.37	30.79	31.20	31.62	32.04	32.45	32.87	33.28	33.70	34.12	34.53	34.95	35.36	35.78
50 – 55%	21.71	22.89	24.08	25.26	26.44	27.63	28.81	29.23	29.64	30.06	30.48	30.89	31.31	31.72	32.14	32.56	32.97	33.39	33.80	34.22	34.64
45 – 50%	20.56	21.75	22.93	24.12	25.30	26.48	27.67	28.08	28.50	28.92	29.33	29.75	30.16	30.58	31.00	31.41	31.83	32.24	32.66	33.08	33.49
40 – 45%	19.42	20.60	21.79	22.97	24.16	25.34	26.52	26.94	27.36	27.77	28.19	28.60	29.02	29.44	29.85	30.27	30.68	31.10	31.52	31.93	32.35
35 – 40%	18.28	19.46	20.64	21.83	23.01	24.20	25.38	25.80	26.21	26.63	27.04	27.46	27.88	28.29	28.71	29.12	29.54	29.96	30.37	30.79	31.20
30 – 35%	17.13	18.32	19.50	20.68	21.87	23.05	24.24	24.65	25.07	25.48	25.90	26.32	26.73	27.15	27.56	27.98	28.40	28.81	29.23	29.64	30.06
25 – 30%	15.99	17.17	18.36	19.54	20.72	21.91	23.09	23.51	23.92	24.34	24.76	25.17	25.59	26.00	26.42	26.84	27.25	27.67	28.08	28.50	28.92
20 – 25%	14.84	16.03	17.21	18.40	19.58	20.76	21.95	22.36	22.78	23.20	23.61	24.03	24.44	24.86	25.28	25.69	26.11	26.52	26.94	27.36	27.77
15 – 20%	13.70	14.88	16.07	17.25	18.44	19.62	20.80	21.22	21.64	22.05	22.47	22.88	23.30	23.72	24.13	24.55	24.96	25.38	25.80	26.21	26.63
10 – 15%	12.56	13.74	14.92	16.11	17.29	18.48	19.66	20.08	20.49	20.91	21.32	21.74	22.16	22.57	22.99	23.40	23.82	24.24	24.65	25.07	25.48
5 – 10%	11.41	12.60	13.78	14.96	16.15	17.33	18.52	18.93	19.35	19.76	20.18	20.60	21.01	21.43	21.84	22.26	22.68	23.09	23.51	23.92	24.34
0 – 5%	10.27	11.45	12.64	13.82	15.00	16.19	17.37	17.79	18.20	18.62	19.04	19.45	19.87	20.28	20.70	21.12	21.53	21.95	22.36	22.78	23.20

FIGURE A.3. The use of tables by an overconfident agent: the 4th guess.

Ihre Schätzung war: 0

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	30.47	31.65	32.84	34.02	35.20	36.39	37.57	38.76	39.94	41.12	42.31	42.72	43.14	43.56	43.97	44.39	44.80	45.22	45.64	46.05	46.47
90 – 95%	29.32	30.51	31.69	32.88	34.06	35.24	36.43	37.61	38.80	39.98	41.16	41.58	42.00	42.41	42.83	43.24	43.66	44.08	44.49	44.91	45.32
85 – 90%	28.18	29.36	30.55	31.73	32.92	34.10	35.28	36.47	37.65	38.84	40.02	40.44	40.85	41.27	41.68	42.10	42.52	42.93	43.35	43.76	44.18
80 – 85%	27.04	28.22	29.40	30.59	31.77	32.96	34.14	35.32	36.51	37.69	38.88	39.29	39.71	40.12	40.54	40.96	41.37	41.79	42.20	42.62	43.04
75 – 80%	25.89	27.08	28.26	29.44	30.63	31.81	33.00	34.18	35.36	36.55	37.73	38.15	38.56	38.98	39.40	39.81	40.23	40.64	41.06	41.48	41.89
70 – 75%	24.75	25.93	27.12	28.30	29.48	30.67	31.85	33.04	34.22	35.40	36.59	37.00	37.42	37.84	38.25	38.67	39.08	39.50	39.92	40.33	40.75
65 – 70%	23.60	24.79	25.97	27.16	28.34	29.52	30.71	31.89	33.08	34.26	35.44	35.86	36.28	36.69	37.11	37.52	37.94	38.36	38.77	39.19	39.60
60 – 65%	22.46	23.64	24.83	26.01	27.20	28.38	29.56	30.75	31.93	33.12	34.30	34.72	35.13	35.55	35.96	36.38	36.80	37.21	37.63	38.04	38.46
55 – 60%	21.32	22.50	23.68	24.87	26.05	27.24	28.42	29.60	30.79	31.97	33.16	33.57	33.99	34.40	34.82	35.24	35.65	36.07	36.48	36.90	37.32
50 – 55%	20.17	21.36	22.54	23.72	24.91	26.09	27.28	28.46	29.64	30.83	32.01	32.43	32.84	33.26	33.68	34.09	34.51	34.92	35.34	35.76	36.17
45 – 50%	19.03	20.21	21.40	22.58	23.76	24.95	26.13	27.32	28.50	29.68	30.87	31.28	31.70	32.12	32.53	32.95	33.36	33.78	34.20	34.61	35.03
40 – 45%	17.88	19.07	20.25	21.44	22.62	23.80	24.99	26.17	27.36	28.54	29.72	30.14	30.56	30.97	31.39	31.80	32.22	32.64	33.05	33.47	33.88
35 – 40%	16.74	17.92	19.11	20.29	21.48	22.66	23.84	25.03	26.21	27.40	28.58	29.00	29.41	29.83	30.24	30.66	31.08	31.49	31.91	32.32	32.74
30 – 35%	15.60	16.78	17.96	19.15	20.33	21.52	22.70	23.88	25.07	26.25	27.44	27.85	28.27	28.68	29.10	29.52	29.93	30.35	30.76	31.18	31.60
25 – 30%	14.45	15.64	16.82	18.00	19.19	20.37	21.56	22.74	23.92	25.11	26.29	26.71	27.12	27.54	27.96	28.37	28.79	29.20	29.62	30.04	30.45
20 – 25%	13.31	14.49	15.68	16.86	18.04	19.23	20.41	21.60	22.78	23.96	25.15	25.56	25.98	26.40	26.81	27.23	27.64	28.06	28.48	28.89	29.31
15 – 20%	12.16	13.35	14.53	15.72	16.90	18.08	19.27	20.45	21.64	22.82	24.00	24.42	24.84	25.25	25.67	26.08	26.50	26.92	27.33	27.75	28.16
10 – 15%	11.02	12.20	13.39	14.57	15.76	16.94	18.12	19.31	20.49	21.68	22.86	23.28	23.69	24.11	24.52	24.94	25.36	25.77	26.19	26.60	27.02
5 – 10%	9.88	11.06	12.24	13.43	14.61	15.80	16.98	18.16	19.35	20.53	21.72	22.13	22.55	22.96	23.38	23.80	24.21	24.63	25.04	25.46	25.88
0 – 5%	8.73	9.92	11.10	12.28	13.47	14.65	15.84	17.02	18.20	19.39	20.57	20.99	21.40	21.82	22.24	22.65	23.07	23.48	23.90	24.32	24.73

FIGURE A.4. The use of tables by an underconfident agent: the 2nd guess.

Ihre Schätzung war: 9

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	27.01	28.20	29.38	30.56	31.75	32.93	34.12	35.30	36.48	37.67	38.85	40.04	41.22	42.40	43.59	44.77	45.96	47.14	48.32	49.51	49.92
90 – 95%	25.87	27.05	28.24	29.42	30.60	31.79	32.97	34.16	35.34	36.52	37.71	38.89	40.08	41.26	42.44	43.63	44.81	46.00	47.18	48.36	48.78
85 – 90%	24.72	25.91	27.09	28.28	29.46	30.64	31.83	33.01	34.20	35.38	36.56	37.75	38.93	40.12	41.30	42.48	43.67	44.85	46.04	47.22	47.64
80 – 85%	23.58	24.76	25.95	27.13	28.32	29.50	30.68	31.87	33.05	34.24	35.42	36.60	37.79	38.97	40.16	41.34	42.52	43.71	44.89	46.08	46.49
75 – 80%	22.44	23.62	24.80	25.99	27.17	28.36	29.54	30.72	31.91	33.09	34.28	35.46	36.64	37.83	39.01	40.20	41.38	42.56	43.75	44.93	45.35
70 – 75%	21.29	22.48	23.66	24.84	26.03	27.21	28.40	29.58	30.76	31.95	33.13	34.32	35.50	36.68	37.87	39.05	40.24	41.42	42.60	43.79	44.20
65 – 70%	20.15	21.33	22.52	23.70	24.88	26.07	27.25	28.44	29.62	30.80	31.99	33.17	34.36	35.54	36.72	37.91	39.09	40.28	41.46	42.64	43.06
60 – 65%	19.00	20.19	21.37	22.56	23.74	24.92	26.11	27.29	28.48	29.66	30.84	32.03	33.21	34.40	35.58	36.76	37.95	39.13	40.32	41.50	41.92
55 – 60%	17.86	19.04	20.23	21.41	22.60	23.78	24.96	26.15	27.33	28.52	29.70	30.88	32.07	33.25	34.44	35.62	36.80	37.99	39.17	40.36	40.77
50 – 55%	16.72	17.90	19.08	20.27	21.45	22.64	23.82	25.00	26.19	27.37	28.56	29.74	30.92	32.11	33.29	34.48	35.66	36.84	38.03	39.21	39.63
45 – 50%	15.57	16.76	17.94	19.12	20.31	21.49	22.68	23.86	25.04	26.23	27.41	28.60	29.78	30.96	32.15	33.33	34.52	35.70	36.88	38.07	38.48
40 – 45%	14.43	15.61	16.80	17.98	19.16	20.35	21.53	22.72	23.90	25.08	26.27	27.45	28.64	29.82	31.00	32.19	33.37	34.56	35.74	36.92	37.34
35 – 40%	13.28	14.47	15.65	16.84	18.02	19.20	20.39	21.57	22.76	23.94	25.12	26.31	27.49	28.68	29.86	31.04	32.23	33.41	34.60	35.78	36.20
30 – 35%	12.14	13.32	14.51	15.69	16.88	18.06	19.24	20.43	21.61	22.80	23.98	25.16	26.35	27.53	28.72	29.90	31.08	32.27	33.45	34.64	35.05
25 – 30%	11.00	12.18	13.36	14.55	15.73	16.92	18.10	19.28	20.47	21.65	22.84	24.02	25.20	26.39	27.57	28.76	29.94	31.12	32.31	33.49	33.91
20 – 25%	9.85	11.04	12.22	13.40	14.59	15.77	16.96	18.14	19.32	20.51	21.69	22.88	24.06	25.24	26.43	27.61	28.80	29.98	31.16	32.35	32.76
15 – 20%	8.71	9.89	11.08	12.26	13.44	14.63	15.81	17.00	18.18	19.36	20.55	21.73	22.92	24.10	25.28	26.47	27.65	28.84	30.02	31.20	31.62
10 – 15%	7.56	8.75	9.93	11.12	12.30	13.48	14.67	15.85	17.04	18.22	19.40	20.59	21.77	22.96	24.14	25.32	26.51	27.69	28.88	30.06	30.48
5 – 10%	6.42	7.60	8.79	9.97	11.16	12.34	13.52	14.71	15.89	17.08	18.26	19.44	20.63	21.81	23.00	24.18	25.36	26.55	27.73	28.92	29.33
0 – 5%	5.28	6.46	7.64	8.83	10.01	11.20	12.38	13.56	14.75	15.93	17.12	18.30	19.48	20.67	21.85	23.04	24.22	25.40	26.59	27.77	28.19

FIGURE A.5. The use of tables by an underconfident agent: the 3rd guess.

Ihre Schätzung war: 6

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	28.16	29.35	30.53	31.72	32.90	34.08	35.27	36.45	37.64	38.82	40.00	41.19	42.37	43.56	44.74	45.92	47.11	47.52	47.94	48.36	48.77
90 – 95%	27.02	28.20	29.39	30.57	31.76	32.94	34.12	35.31	36.49	37.68	38.86	40.04	41.23	42.41	43.60	44.78	45.96	46.38	46.80	47.21	47.63
85 – 90%	25.88	27.06	28.24	29.43	30.61	31.80	32.98	34.16	35.35	36.53	37.72	38.90	40.08	41.27	42.45	43.64	44.82	45.24	45.65	46.07	46.48
80 – 85%	24.73	25.92	27.10	28.28	29.47	30.65	31.84	33.02	34.20	35.39	36.57	37.76	38.94	40.12	41.31	42.49	43.68	44.09	44.51	44.92	45.34
75 – 80%	23.59	24.77	25.96	27.14	28.32	29.51	30.69	31.88	33.06	34.24	35.43	36.61	37.80	38.98	40.16	41.35	42.53	42.95	43.36	43.78	44.20
70 – 75%	22.44	23.63	24.81	26.00	27.18	28.36	29.55	30.73	31.92	33.10	34.28	35.47	36.65	37.84	39.02	40.20	41.39	41.80	42.22	42.64	43.05
65 – 70%	21.30	22.48	23.67	24.85	26.04	27.22	28.40	29.59	30.77	31.96	33.14	34.32	35.51	36.69	37.88	39.06	40.24	40.66	41.08	41.49	41.91
60 – 65%	20.16	21.34	22.52	23.71	24.89	26.08	27.26	28.44	29.63	30.81	32.00	33.18	34.36	35.55	36.73	37.92	39.10	39.52	39.93	40.35	40.76
55 – 60%	19.01	20.20	21.38	22.56	23.75	24.93	26.12	27.30	28.48	29.67	30.85	32.04	33.22	34.40	35.59	36.77	37.96	38.37	38.79	39.20	39.62
50 – 55%	17.87	19.05	20.24	21.42	22.60	23.79	24.97	26.16	27.34	28.52	29.71	30.89	32.08	33.26	34.44	35.63	36.81	37.23	37.64	38.06	38.48
45 – 50%	16.72	17.91	19.09	20.28	21.46	22.64	23.83	25.01	26.20	27.38	28.56	29.75	30.93	32.12	33.30	34.48	35.67	36.08	36.50	36.92	37.33
40 – 45%	15.58	16.76	17.95	19.13	20.32	21.50	22.68	23.87	25.05	26.24	27.42	28.60	29.79	30.97	32.16	33.34	34.52	34.94	35.36	35.77	36.19
35 – 40%	14.44	15.62	16.80	17.99	19.17	20.36	21.54	22.72	23.91	25.09	26.28	27.46	28.64	29.83	31.01	32.20	33.38	33.80	34.21	34.63	35.04
30 – 35%	13.29	14.48	15.66	16.84	18.03	19.21	20.40	21.58	22.76	23.95	25.13	26.32	27.50	28.68	29.87	31.05	32.24	32.65	33.07	33.48	33.90
25 – 30%	12.15	13.33	14.52	15.70	16.88	18.07	19.25	20.44	21.62	22.80	23.99	25.17	26.36	27.54	28.72	29.91	31.09	31.51	31.92	32.34	32.76
20 – 25%	11.00	12.19	13.37	14.56	15.74	16.92	18.11	19.29	20.48	21.66	22.84	24.03	25.21	26.40	27.58	28.76	29.95	30.36	30.78	31.20	31.61
15 – 20%	9.86	11.04	12.23	13.41	14.60	15.78	16.96	18.15	19.33	20.52	21.70	22.88	24.07	25.25	26.44	27.62	28.80	29.22	29.64	30.05	30.47
10 – 15%	8.72	9.90	11.08	12.27	13.45	14.64	15.82	17.00	18.19	19.37	20.56	21.74	22.92	24.11	25.29	26.48	27.66	28.08	28.49	28.91	29.32
5 – 10%	7.57	8.76	9.94	11.12	12.31	13.49	14.68	15.86	17.04	18.23	19.41	20.60	21.78	22.96	24.15	25.33	26.52	26.93	27.35	27.76	28.18
0 – 5%	6.43	7.61	8.80	9.98	11.16	12.35	13.53	14.72	15.90	17.08	18.27	19.45	20.64	21.82	23.00	24.19	25.37	25.79	26.20	26.62	27.04

FIGURE A.6. The use of tables by an underconfident agent: the 4th guess.

Ihre Schätzung war: 0

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	30.47	31.65	32.84	34.02	35.20	36.39	37.57	38.76	39.94	41.12	42.31	42.72	43.14	43.56	43.97	44.39	44.80	45.22	45.64	46.05	46.47
90 – 95%	29.32	30.51	31.69	32.88	34.06	35.24	36.43	37.61	38.80	39.98	41.16	41.58	42.00	42.41	42.83	43.24	43.66	44.08	44.49	44.91	45.32
85 – 90%	28.18	29.36	30.55	31.73	32.92	34.10	35.28	36.47	37.65	38.84	40.02	40.44	40.85	41.27	41.68	42.10	42.52	42.93	43.35	43.76	44.18
80 – 85%	27.04	28.22	29.40	30.59	31.77	32.96	34.14	35.32	36.51	37.69	38.88	39.29	39.71	40.12	40.54	40.96	41.37	41.79	42.20	42.62	43.04
75 – 80%	25.89	27.08	28.26	29.44	30.63	31.81	33.00	34.18	35.36	36.55	37.73	38.15	38.56	38.98	39.40	39.81	40.23	40.64	41.06	41.48	41.89
70 – 75%	24.75	25.93	27.12	28.30	29.48	30.67	31.85	33.04	34.22	35.40	36.59	37.00	37.42	37.84	38.25	38.67	39.08	39.50	39.92	40.33	40.75
65 – 70%	23.60	24.79	25.97	27.16	28.34	29.52	30.71	31.89	33.08	34.26	35.44	35.86	36.28	36.69	37.11	37.52	37.94	38.36	38.77	39.19	39.60
60 – 65%	22.46	23.64	24.83	26.01	27.20	28.38	29.56	30.75	31.93	33.12	34.30	34.72	35.13	35.55	35.96	36.38	36.80	37.21	37.63	38.04	38.46
55 – 60%	21.32	22.50	23.68	24.87	26.05	27.24	28.42	29.60	30.79	31.97	33.16	33.57	33.99	34.40	34.82	35.24	35.65	36.07	36.48	36.90	37.32
50 – 55%	20.17	21.36	22.54	23.72	24.91	26.09	27.28	28.46	29.64	30.83	32.01	32.43	32.84	33.26	33.68	34.09	34.51	34.92	35.34	35.76	36.17
45 – 50%	19.03	20.21	21.40	22.58	23.76	24.95	26.13	27.32	28.50	29.68	30.87	31.28	31.70	32.12	32.53	32.95	33.36	33.78	34.20	34.61	35.03
40 – 45%	17.88	19.07	20.25	21.44	22.62	23.80	24.99	26.17	27.36	28.54	29.72	30.14	30.56	30.97	31.39	31.80	32.22	32.64	33.05	33.47	33.88
35 – 40%	16.74	17.92	19.11	20.29	21.48	22.66	23.84	25.03	26.21	27.40	28.58	29.00	29.41	29.83	30.24	30.66	31.08	31.49	31.91	32.32	32.74
30 – 35%	15.60	16.78	17.96	19.15	20.33	21.52	22.70	23.88	25.07	26.25	27.44	27.85	28.27	28.68	29.10	29.52	29.93	30.35	30.76	31.18	31.60
25 – 30%	14.45	15.64	16.82	18.00	19.19	20.37	21.56	22.74	23.92	25.11	26.29	26.71	27.12	27.54	27.96	28.37	28.79	29.20	29.62	30.04	30.45
20 – 25%	13.31	14.49	15.68	16.86	18.04	19.23	20.41	21.60	22.78	23.96	25.15	25.56	25.98	26.40	26.81	27.23	27.64	28.06	28.48	28.89	29.31
15 – 20%	12.16	13.35	14.53	15.72	16.90	18.08	19.27	20.45	21.64	22.82	24.00	24.42	24.84	25.25	25.67	26.08	26.50	26.92	27.33	27.75	28.16
10 – 15%	11.02	12.20	13.39	14.57	15.76	16.94	18.12	19.31	20.49	21.68	22.86	23.28	23.69	24.11	24.52	24.94	25.36	25.77	26.19	26.60	27.02
5 – 10%	9.88	11.06	12.24	13.43	14.61	15.80	16.98	18.16	19.35	20.53	21.72	22.13	22.55	22.96	23.38	23.80	24.21	24.63	25.04	25.46	25.88
0 – 5%	8.73	9.92	11.10	12.28	13.47	14.65	15.84	17.02	18.20	19.39	20.57	20.99	21.40	21.82	22.24	22.65	23.07	23.48	23.90	24.32	24.73

FIGURE A.7. The use of tables by an unbiased agent: the 2nd guess.

Ihre Schätzung war: −4

Leistungsintervall	Mögliche Zufallzahl:																				
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
95 – 100%	32.00	33.19	34.37	35.56	36.74	37.92	39.11	39.52	39.94	40.36	40.77	41.19	41.60	42.02	42.44	42.85	43.27	43.68	44.10	44.52	44.93
90 – 95%	30.86	32.04	33.23	34.41	35.60	36.78	37.96	38.38	38.80	39.21	39.63	40.04	40.46	40.88	41.29	41.71	42.12	42.54	42.96	43.37	43.79
85 – 90%	29.72	30.90	32.08	33.27	34.45	35.64	36.82	37.24	37.65	38.07	38.48	38.90	39.32	39.73	40.15	40.56	40.98	41.40	41.81	42.23	42.64
80 – 85%	28.57	29.76	30.94	32.12	33.31	34.49	35.68	36.09	36.51	36.92	37.34	37.76	38.17	38.59	39.00	39.42	39.84	40.25	40.67	41.08	41.50
75 – 80%	27.43	28.61	29.80	30.98	32.16	33.35	34.53	34.95	35.36	35.78	36.20	36.61	37.03	37.44	37.86	38.28	38.69	39.11	39.52	39.94	40.36
70 – 75%	26.28	27.47	28.65	29.84	31.02	32.20	33.39	33.80	34.22	34.64	35.05	35.47	35.88	36.30	36.72	37.13	37.55	37.96	38.38	38.80	39.21
65 – 70%	25.14	26.32	27.51	28.69	29.88	31.06	32.24	32.66	33.08	33.49	33.91	34.32	34.74	35.16	35.57	35.99	36.40	36.82	37.24	37.65	38.07
60 – 65%	24.00	25.18	26.36	27.55	28.73	29.92	31.10	31.52	31.93	32.35	32.76	33.18	33.60	34.01	34.43	34.84	35.26	35.68	36.09	36.51	36.92
55 – 60%	22.85	24.04	25.22	26.40	27.59	28.77	29.96	30.37	30.79	31.20	31.62	32.04	32.45	32.87	33.28	33.70	34.12	34.53	34.95	35.36	35.78
50 – 55%	21.71	22.89	24.08	25.26	26.44	27.63	28.81	29.23	29.64	30.06	30.48	30.89	31.31	31.72	32.14	32.56	32.97	33.39	33.80	34.22	34.64
45 – 50%	20.56	21.75	22.93	24.12	25.30	26.48	27.67	28.08	28.50	28.92	29.33	29.75	30.16	30.58	31.00	31.41	31.83	32.24	32.66	33.08	33.49
40 – 45%	19.42	20.60	21.79	22.97	24.16	25.34	26.52	26.94	27.36	27.77	28.19	28.60	29.02	29.44	29.85	30.27	30.68	31.10	31.52	31.93	32.35
35 – 40%	18.28	19.46	20.64	21.83	23.01	24.20	25.38	25.80	26.21	26.63	27.04	27.46	27.88	28.29	28.71	29.12	29.54	29.96	30.37	30.79	31.20
30 – 35%	17.13	18.32	19.50	20.68	21.87	23.05	24.24	24.65	25.07	25.48	25.90	26.32	26.73	27.15	27.56	27.98	28.40	28.81	29.23	29.64	30.06
25 – 30%	15.99	17.17	18.36	19.54	20.72	21.91	23.09	23.51	23.92	24.34	24.76	25.17	25.59	26.00	26.42	26.84	27.25	27.67	28.08	28.50	28.92
20 – 25%	14.84	16.03	17.21	18.40	19.58	20.76	21.95	22.36	22.78	23.20	23.61	24.03	24.44	24.86	25.28	25.69	26.11	26.52	26.94	27.36	27.77
15 – 20%	13.70	14.88	16.07	17.25	18.44	19.62	20.80	21.22	21.64	22.05	22.47	22.88	23.30	23.72	24.13	24.55	24.96	25.38	25.80	26.21	26.63
10 – 15%	12.56	13.74	14.92	16.11	17.29	18.48	19.66	20.08	20.49	20.91	21.32	21.74	22.16	22.57	22.99	23.40	23.82	24.24	24.65	25.07	25.48
5 – 10%	11.41	12.60	13.78	14.96	16.15	17.33	18.52	18.93	19.35	19.76	20.18	20.60	21.01	21.43	21.84	22.26	22.68	23.09	23.51	23.92	24.34
0 – 5%	10.27	11.45	12.64	13.82	15.00	16.19	17.37	17.79	18.20	18.62	19.04	19.45	19.87	20.28	20.70	21.12	21.53	21.95	22.36	22.78	23.20

FIGURE A.8. The use of tables by an unbiased agent: the 3rd and the 4th guess.

APPENDIX B

Misguided Learning: Additional Results

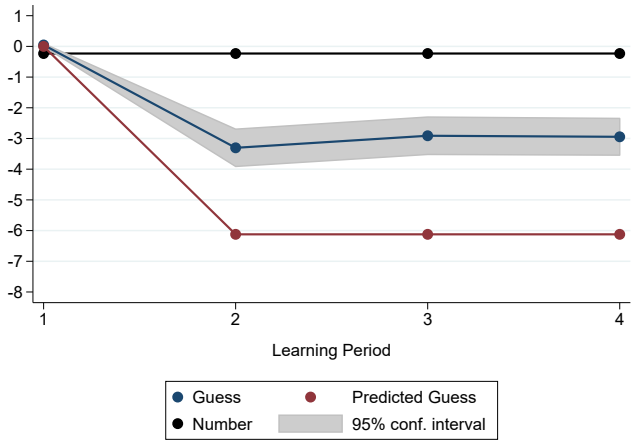
In this section, we present results complementing Section 1.3.2 of the paper. We describe decisions in the single-feedback rounds for the three types of agents in Section B.1. In Section B.2, we gather the estimates based on the pooled sample (described in the last paragraph in Section 1.3.2.3). In Section B.3, we present tables complementing Table 1.5 from the paper. Lastly, we present evidence on the model's performance.

B.1. The single-feedback rounds

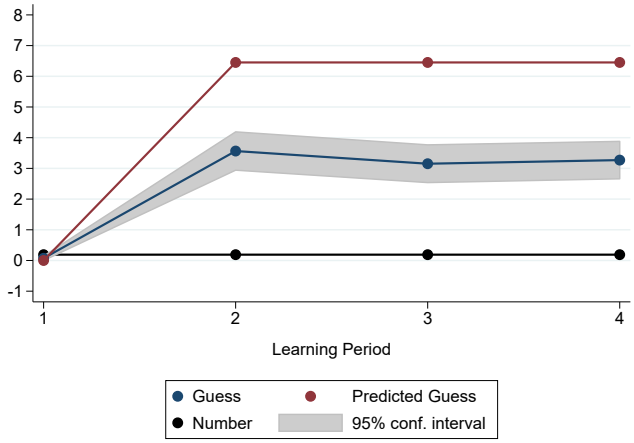
We present graphically the decisions of overconfident, underconfident, and unbiased agents in the single-feedback rounds. Figure B.1 corresponds to Figures 1.3 and 1.2 in the paper. Recall that, in the single-feedback rounds, feedback was independent of subjects' guesses (participants were aware that the number displayed after the 2^{nd} and the 3^{rd} guess will be based on their 1^{st} guess). Thus, there is no reason for subjects to change their decisions – the predicted 2^{nd}, 3^{rd}, and 4^{th} guess are of the same value.

In Table B.1, we present the results of comparing pairs of coefficients from regressions in Tables 1.3 and 1.4 in the paper. The tests are described in Sections 1.3.2.1 and 1.3.2.2 in the paper. Here, we only gather them in one table for completeness.

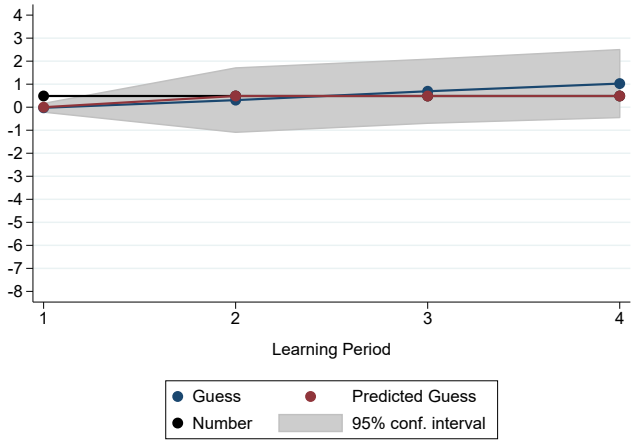
FIGURE B.1. Learning process in the single-feedback rounds.



(A) Overconfident agents in SF Rounds.



(B) Underconfident agents in SF Rounds.



(C) Unbiased agents in SF Rounds.

TABLE B.1. The regression coefficients in the multiple- and single-feedback rounds in the ego-relevant condition.

(a) Overconfident Agents

	H_0: $\beta^2_{MF} \leq \beta^3_{MF}$	H_0: $\beta^3_{MF} \leq \beta^4_{MF}$	H_0: $\beta^2_{MF} \leq \beta^4_{MF}$
p-value	0.019^{**}	0.159	0.003^{***}
	H_0: $\beta^2_{SF} \leq \beta^3_{SF}$	H_0: $\beta^3_{SF} \leq \beta^4_{SF}$	H_0: $\beta^2_{SF} \leq \beta^4_{SF}$
p-value	0.953	0.431	0.958

(b) Unbiased Agents

	H_0: $\beta^2_{MF} = \beta^3_{MF}$	H_0: $\beta^3_{MF} = \beta^4_{MF}$	H_0: $\beta^2_{MF} = \beta^4_{MF}$
p-value	0.056^{*}	0.885	0.102
	H_0: $\beta^2_{SF} = \beta^3_{SF}$	H_0: $\beta^3_{SF} = \beta^4_{SF}$	H_0: $\beta^2_{SF} = \beta^4_{SF}$
p-value	0.251	0.307	0.226

(c) Underconfident Agents

	H_0: $\beta^2_{MF} \leq \beta^3_{MF}$	H_0: $\beta^3_{MF} = \beta^4_{MF}$
p-value	0.000^{***}	0.681
	H_0: $\beta^2_{SF} \leq \beta^3_{SF}$	H_0: $\beta^3_{SF} = \beta^4_{SF}$
p-value	0.008^{***}	0.394

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

B.2. The effect of providing informative feedback

In this section, we present the analysis based on pooled data from the multiple- and single-feedback rounds. We look at the effect of receiving informative feedback (the "MF Round" variable) on learning. In the specification presented in Table B.2, the dependent variable is the difference between a subject's guess and the number. The results for overconfident agents are described in the last paragraph in Section 1.3.2.3. For underconfident agents, receiving informative feedback reduces the difference between a guess and the number by 1.29 in the 3^{rd} guess (one-tailed test: p-value = 0.000), and by 1.47 in the 4^{th} guess (one-tailed test: p-value = 0.000). The direction of the effect is in line with the model predictions.[1] As expected, informative feedback does not affect unbiased subjects. In another specification, presented in Table B.3, we use the absolute difference between a guess and the number as a dependent variable.[2] Because of the absolute value, the effect in the second specification is positive for overconfident agents (informative feedback enlarges the absolute difference). Taking this into account, one can conclude that the two specifications yield consistent results.

In the specification presented in Table B.4, the dependent variable is the difference between the 4^{th} and the 2^{nd} guess. We look at participants' decisions after the 2^{nd} guess, because only at this point the two conditions diverge (after the 1^{st} guess, participants received informative feedback both in the multiple- and single-feedback rounds). We interpret the difference between the 4^{th} and the 2^{nd} guess as a change in beliefs about the number in the final guesses. As it is evident in Table B.4, being in a multiple-feedback round makes overconfident participants more pessimistic about the number by around -1.19, which constitutes 67% of the effect predicted by the model. In the case of underconfident agents, the coefficient captures the degree of correction after the second guess. It is equal to -1.23 (68% of the effect predicted by the model) and significant at the 1%-level. Taken together, the results support our claim that the effect is driven by the model's mechanism and not by external factors.

---

[1] The model predicts that in the 3^{rd} guess underconfident agents correct their decisions from the 2^{nd} guess. In the single-feedback rounds, however, this is no longer the case, as agents do not receive any meaningful feedback after the 2^{nd} guess. Consequently, the effect of being in a multiple-feedback round is negative – the sign indicates the correction after the second feedback.

[2] Although this specification might be viewed as more appropriate, we decided to include the other one in the main text because it can be directly linked to the graphs and the sign of the effect is indicative of agents' pessimism (optimism) about the number.

TABLE B.2. The effect of feedback on difference between guess and number.

	Overconfident	Unbiased Agents	Underconfident
	(1)	(2)	(3)
Dependent variable: the difference between the number and the 4^{th} guess.			
MF Round	-1.810***	-0.128	-1.468***
	(0.391)	(0.422)	(0.268)
Const.	-1.264**	0.538	2.229***
	(0.545)	(0.408)	(0.526)
Dependent variable: the difference between the number and the 3^{rd} guess.			
MF Round	-1.570***	0.256	-1.291***
	(0.363)	(0.215)	(0.287)
Const.	-1.230**	0.205	1.898***
	(0.531)	(0.154)	(0.500)
Dependent variable: the difference between the number and the 2^{nd} guess.			
MF Round	-0.616*	0.051	-0.236
	(0.334)	(0.214)	(0.259)
Const.	-1.207**	-0.179	2.625***
	(0.550)	(0.185)	(0.558)
N	474	78	474

"MF Round" is a dummy variable taking value 1 if in a multiple-feedback round.
Controlling for subjects' relative performance does not change the results.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B.3. The effect of feedback on absolute difference between guess and number.

	Overconfident	Unbiased Agents	Underconfident
	(1)	(2)	(3)
Dependent variable: the absolute difference between the number and the 4^{th} guess.			
MF Round	1.211***	-0.333	-1.308***
	(0.303)	(0.423)	(0.235)
Const.	1.895***	0.949	2.924***
	(0.433)	(0.512)	(0.438)
Dependent variable: the absolute difference between the number and the 3^{rd} guess.			
MF Round	1.122***	0.0513	-1.350***
	(0.293)	(0.268)	(0.230)
Const.	1.692***	0.615**	2.737***
	(0.443)	(0.197)	(0.408)
Dependent variable: the absolute difference between the number and the 2^{nd} guess.			
MF Round	0.236	-0.205	-0.0928
	(0.221)	(0.206)	(0.212)
Const.	1.494***	0.333	3.086***
	(0.361)	(0.197)	(0.453)
N	474	78	474

"MF Round" is a dummy variable taking value 1 if in a multiple-feedback round. Controlling for subjects' relative performance does not change the results.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B.4. The effect of informative feedback on learning.

	Dependent variable: the difference between the 4^{th} and the 2^{nd} guess.		
	Overconfident	*Unbiased*	*Underconfident*
	(1)	(2)	(3)
MF Round	-1.194***	-0.179	-1.232***
	(0.337)	(0.499)	(0.267)
Const.	0.359*	0.718	-0.295*
	(0.205)	(0.559)	(0.153)
Observations	474	78	474

Note: The dependent variable is the difference between the 4^{th} and the 2^{nd} guess. The independent variable "MF Round" is a dummy variable taking value 1 if the round is a multiple-feedback round. Controlling for the number being guessed and subjects' relative performance does not change the results.

Standard errors clustered at the individual level. Their values in parentheses.

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

B.3. The effect of initial bias

In Table B.5, we present the estimation results from Table 1.5 in the paper based on a sample of underconfident and overconfident agents excluding unbiased participants. While the coefficients are quantitatively different from the one in Table 1.5 in the paper, the direction of the effect remains the same. In Table B.6, we gather corresponding results based on the data from the single-feedback rounds.

TABLE B.5. The effect of bias in MF rounds.

	Overconfident (1)		Underconfident (2)	
Dependent variable: the difference between a guess and the number in MF rounds. Independent variables: dummy variables for each guess and their interactions.				
2^{nd} guess MF	-2.162***	(0.608)	2.257***	(0.717)
3^{rd} guess MF	-2.949***	(0.645)	0.302	(0.557)
4^{th} guess MF	-3.248***	(0.713)	0.259	(0.597)
Bias	-0.782	(1.318)	-1.720	(1.798)
Bias × 2^{nd} guess MF	-5.191**	(2.495)	-2.654	(3.139)
Bias × 3^{rd} guess MF	-4.421*	(2.344)	-5.064**	(2.297)
Bias × 4^{th} guess MF	-4.337*	(2.608)	-4.985**	(2.448)
Const.	0.225	(0.454)	0.001	(0.498)
N	948		948	

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE B.6. The effect of bias in SF rounds.

	Overconfident or Unbiased (1)		Underconfident or Unbiased (2)	
Dependent variable: the difference between a guess and the number in SF rounds. Independent variables: dummy variables for each guess and their interactions.				
2nd guess SF	-0.980*	(0.516)	1.761***	(0.562)
3rd guess SF	-0.960*	(0.502)	1.569***	(0.533)
4th guess SF	-0.836	(0.515)	2.022***	(0.544)
Bias	-0.305	(1.287)	0.215	(1.739)
Bias × 2nd guess SF	-7.351***	(2.169)	-7.418**	(2.831)
Bias × 3rd guess SF	-5.876***	(2.216)	-6.790***	(2.525)
Bias × 4th guess SF	-6.295***	(2.050)	-5.035*	(2.623)
Const.	0.243	(0.420)	-0.137	(0.384)
N	1104		1104	
	Overconfident (1)		Underconfident (2)	
Dependent variable: the difference between a guess and the number in SF rounds. Independent variables: dummy variables for each guess and their interactions.				
2nd guess SF	-1.821***	(0.661)	2.658***	(0.732)
3rd guess SF	-2.033***	(0.613)	2.104***	(0.708)
4th guess SF	-2.044***	(0.593)	2.632***	(0.707)
Bias	-1.531	(1.426)	1.079	(1.850)
Bias × 2nd guess SF	-5.220**	(2.544)	-4.138	(3.274)
Bias × 3rd guess SF	-3.155	(2.549)	-4.834	(2.999)
Bias × 4th guess SF	-3.234	(2.299)	-2.805	(3.054)
Const.	0.727	(0.497)	0.100	(0.428)
N	948		948	

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B.4. Model's performance

In this section we address the question of the model's explanatory power. We test how well the model explains our data and report the results in Tables B.7, B.8, and B.9. Firstly, we pool the data from the multiple- and single-feedback rounds and look separately at early and late rounds. We refer to the first three rounds as "early rounds", and to the last three rounds as "late rounds". Secondly, we distinguish overconfident, underconfident, and unbiased agents; we look at the model's performance in the groups.

The model seems to better explain the data in early rounds (especially in the first round) than in later rounds. The results are in line with the observation that, during the experiment, subjects were updating their beliefs about their relative ability. At early stages of the experiment, subjects' beliefs were closer to those assumed in the model. The estimation results gathered in Table B.9 demonstrate that choices of the unbiased agents are well-explained by the model. With the R^2 of 0.85 the model explains much variation in the data. The fit is less adequate in case of underconfident agents and much worse for overconfident subjects.

TABLE B.7. Model's performance in early and late rounds.

	All Rounds	Early Rounds	Late Rounds	1^{st} Round
	(1)	(2)	(3)	(4)
Model	0.563***	0.633***	0.493***	0.688***
	(0.030)	(0.031)	(0.036)	(0.035)
Const.	-0.119	-0.102	-0.132	-0.0213
	(0.182)	(0.185)	(0.217)	(0.213)
R^2	0.523	0.605	0.441	0.696
N	3078	1539	1539	513

The dependent variable denotes subjects' actual guesses. The independent variable "Model" denotes guesses predicted by the model.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B.8. Model's performance in multiple- and single-feedback rounds.

	All Rounds		Early Rounds		Late Rounds	
	(SF)	(MF)	(SF)	(MF)	(SF)	(MF)
Model	0.559***	0.563***	0.609***	0.650***	0.502***	0.482***
	(0.034)	(0.032)	(0.040)	(0.033)	(0.043)	(0.042)
Const.	0.0731	-0.310	0.150	-0.340	-0.0499	-0.213
	(0.212)	(0.181)	(0.249)	(0.193)	(0.268)	(0.239)
R^2	0.516	0.522	0.567	0.634	0.458	0.422
N	1539	1539	813	726	726	813

The dependent variable denotes subjects' actual guesses. The independent variable "Model" denotes guesses predicted by the model.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B.9. Model's performance for different types of agents.

	Overconfident	Unbiased Agents	Underconfident
Model	0.575***	0.969***	0.689***
	(0.068)	(0.028)	(0.035)
Const.	0.247	0.220	-1.151***
	(0.312)	(0.132)	(0.220)
R^2	0.182	0.850	0.463
N	1422	234	1422

The dependent variable denotes subjects' actual guesses. The independent variable "Model" denotes guesses predicted by the model.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APPENDIX C

Revealed Beliefs

C.1. Deriving beliefs from guesses

The data from Confidence I and II tells us little about the changes in subjects' beliefs about their performance *during* the learning exercise. To investigate this issue, we attempt to retrieve agents' beliefs from their guesses. The experimental design enables us to divulge the beliefs about one's relative performance with few additional assumptions. The loss-function specification implies that the myopically optimal action is to enter one's beliefs about the number in every guess. There is only one ability level that "rationalizes" the agent's optimal guess, given the feedback he obtained. Thus, to derive agents' beliefs from their actions, we need to assume that the participants chose optimally in every period and without errors.

Assumption R1. (Optimal Actions)

The agent chooses his action optimally and without mistakes in every period.

In every round, we can derive beliefs about the relative performance parameter from the 2^{nd}, the 3^{rd} and the 4^{th} guess. In principle, we can use all 18 revealed beliefs to examine beliefs formation during the task. However, we decided to use only beliefs revealed from the second guess in each round to obtain a less noisy measure (agents might make more mistakes or start experimenting in later trials).

Assumption R2. (Updating at the beginning of the round)

The agents updates beliefs about his performance right before the second guess each round and keeps them unchanged till the beginning of the next round. In other words, the second guess in each round reveals the agent's beliefs in that round.

C.2. Beliefs revealed in rounds 1 to 6

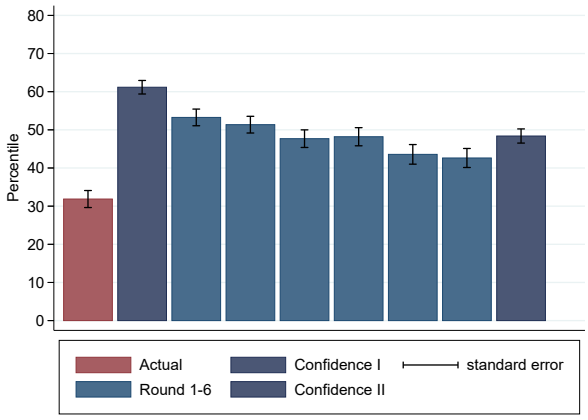
It is instructive to juxtapose the revealed beliefs with the beliefs elicited before and after the learning exercise. In Figure C.1, we present the mean relative performance, beliefs elicited in Confidence I and Confidence II, and beliefs retrieved from the 2nd guess in each round. The beliefs derived from agents' guesses seem to be consistent with the beliefs elicited before and after the learning exercise.[1] From the first to the last round, we observe a gradual change in beliefs in the direction of the true performance level for the overconfident and underconfident agents. The cumulative effect of updating over rounds, measured as the difference between beliefs revealed in the first and last round, is significant for the overconfident and underconfident, but not for the unbiased agents.

To describe the revealed beliefs, complementing the data discussed so far, we present the distributions of beliefs in terms of subjects' bias. In Figure C.2, we present the distribution of bias based on the beliefs elicited in Confidence I and II in panels (a) and (h), and the bias based on the beliefs revealed in rounds 1 to 6 in panels (b) to (g). There is a notable heterogeneity among participants with respect to the magnitude of bias. The distribution changes visibly from round to round, with more participants becoming unbiased towards the end of the experiment. Neither the distributions presented in panels (a) and (b), nor the distributions shown in (g) and (h), are alike.[2] It might be due to the differences in the two elicitation methods or the feedback provided to the subjects (see footnote 2). In particular, the feedback provided after the 1st guess is likely to have a large effect on beliefs revealed in the first round.

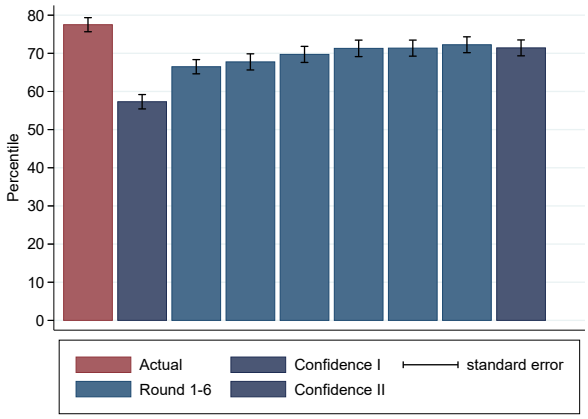
[1] When comparing the elicited and revealed beliefs one should keep in mind several points. Firstly, between Confidence I and the 2nd guess in Round 1, as well as the 2nd guess in Round 6 and Confidence II, agents received feedback that was likely to change their beliefs. Secondly, the two elicitation methods are very different, and participants may not be invariant to the two procedures.

[2] Looking at the last two panels, one can notice that over 35% of all participants entered their choices in Round 6 as if they were unbiased, but only 25% indicated their actual performance as a switching probability in Confidence II. We suspect that the difference is due to dissimilar elicitation methods or agents' (unwarranted) attempt to hedge, rather than participants "unlearning" their abilities at the end of the last round.

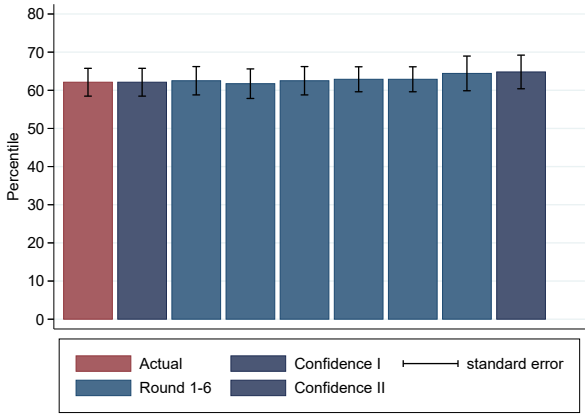
FIGURE C.1. The average performance and beliefs.



(A) Overconfident Agents

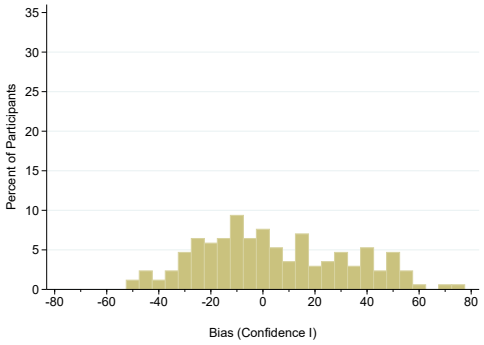


(B) Underconfident Agents

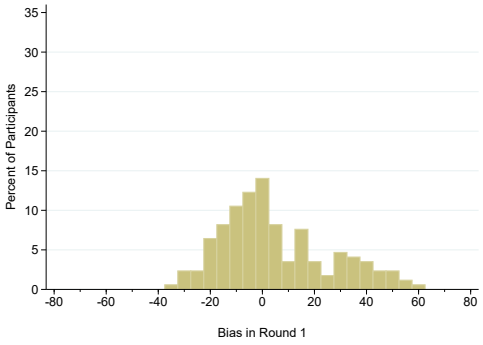


(C) Unbiased Agents

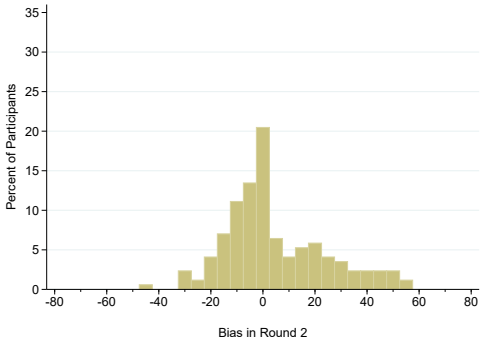
FIGURE C.2. Distribution of participants' bias.



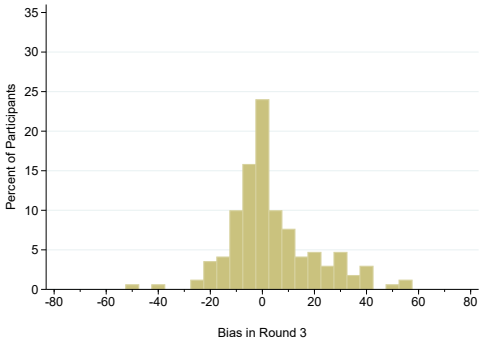
(A) Bias elicited in Confidence I.



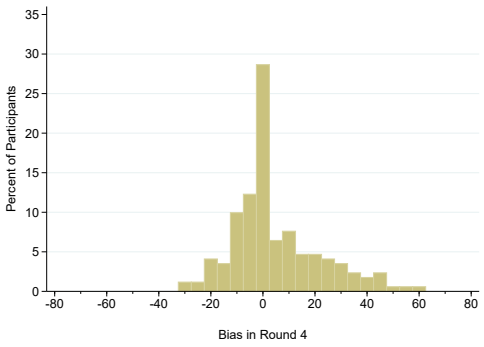
(B) Bias revealed in Round 1.



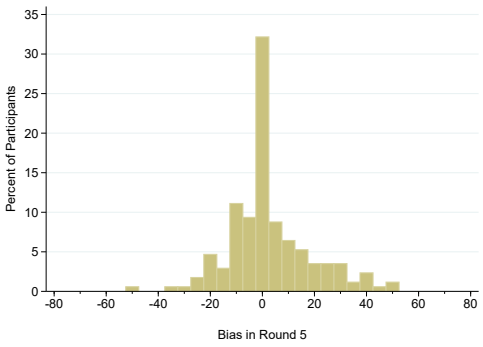
(C) Bias revealed in Round 2.



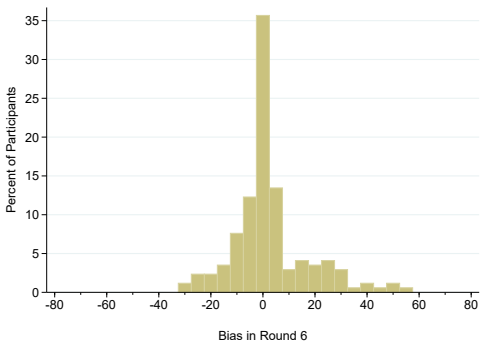
(D) Bias revealed in Round 3.



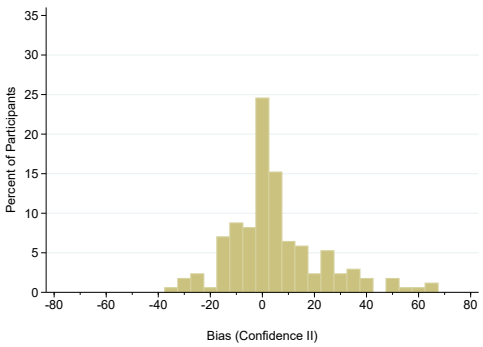
(E) Bias revealed in Round 4.



(F) Bias revealed in Round 5.



(G) Bias revealed in Round 6.



(H) Bias elicited Confidence II.

TABLE C.1. The number of subjects of different type classified as unbiased.

	Conf. I	R1	R2	R3	R4	R5	R6	**Conf. II**
Overconfident*	0	9	7	9	10	14	17	14
Unbiased	13	12	12	12	12	12	12	4
Underconfident	0	3	16	20	27	29	32	24
All subjects	13	24	35	41	49	55	61	42

* Classification based on Confidence I.

Table C.1 presents the number of participants becoming unbiased during the course of the experiment based on beliefs elicited (Conf. I and Conf. II), and revealed (R1 to R6). The agents classified as underconfident in Confidence I are more likely to become unbiased during the experiment than the overconfident agents. 32 participants out of 79 classified as underconfident entered their guesses in the sixth round as if they were unbiased, but only 17 out of 79 overconfident agents did so. Almost all agents classified as unbiased in Confidence I entered their choices as if they were unbiased, but only one third of them indicated the switching point equal to their relative performance in Confidence II. We can only speculate whether the agents were driven by an impulse to hedge, or encountering no difficulties during the main task served as some kind of a signal.

C.3. Model predictions based on revealed beliefs

So far, we have tested the model's predictions assuming that there is no change in agents' beliefs during the experiment. We relax this assumption here, allowing agents to update their beliefs at the beginning of each round. For each agent, we calculate the predicted actions based on his revealed beliefs.

In Figure C.3, we plot the average actual guess and the average guess predicted by the model, separately for the overconfident, underconfident and unbiased agents in the multiple- and single-feedback rounds. Compared to the model predictions based on elicited beliefs, the average predicted guesses (in red) are much closer to the actual choices (in blue). The better fit is reflected in the regression estimates in Table C.2. The coefficients of the Model variable are higher than the respective coefficients in Tables B.7 and B.8 in the previous section, and now there is little difference between the early and late rounds. Overall, the model explains 73.5% variation in the data. Moreover, it does a much better job at explaining the

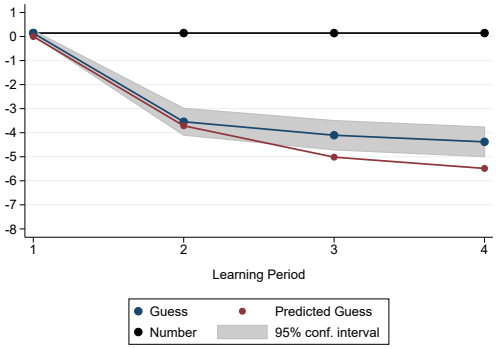
choices of overconfident and underconfident agents, in comparison to the analysis based on elicited beliefs.

Secondly, we re-examine the impact of agent's bias on learning. To this end, we look at the distance between the agent's guess and the number. We classify participants as overconfident, underconfident or unbiased on the basis of their revealed beliefs.[3] In Table C.2, we gathered the estimates for subjects' guesses in the multiple-feedback rounds. Comparing with the results based on elicited beliefs (see Table 1.5 in the paper), the effect of subjects' bias is much stronger. For overconfident agents, subjects' guesses are no longer significant unless interacted with participant's bias. It should not come as a surprise, since the main mechanism of the model operates through the agent's bias. Using a more accurate measure of beliefs leads to a higher and more precise estimates of the effect of subjects' bias.

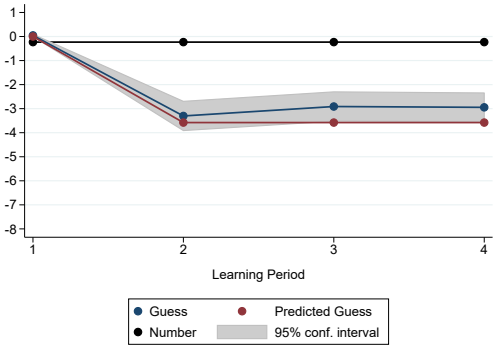
The results presented in this section lend further support to the claim that the differences between theoretical predictions based on elicited beliefs and the actual guesses are due to participants learning about their ability during the task. If we use an alternative measure of beliefs, allowing for updating from round to round, the model closely tracts subjects' behavior.

[3]It is possible for an agent to change his type at the beginning of a round. For this reason, the groups of overconfident and underconfident agents are no longer equinumerous.

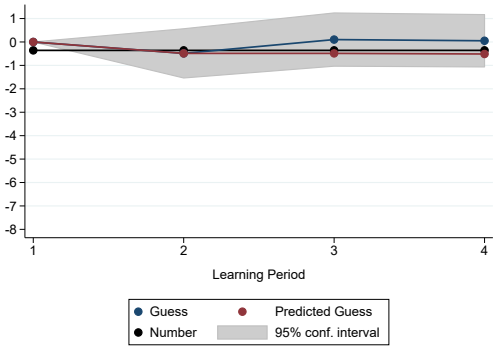
FIGURE C.3. The estimated numbers, the participants' actual and predicted guesses.



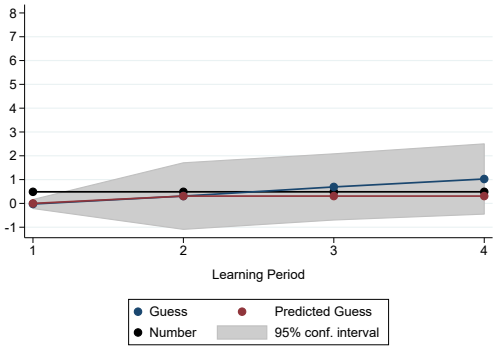
(A) Overconfident in MF rounds.



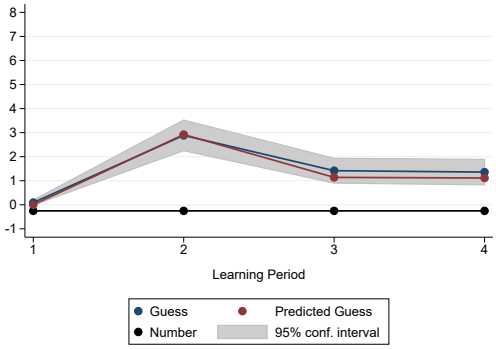
(B) Overconfident in SF rounds.



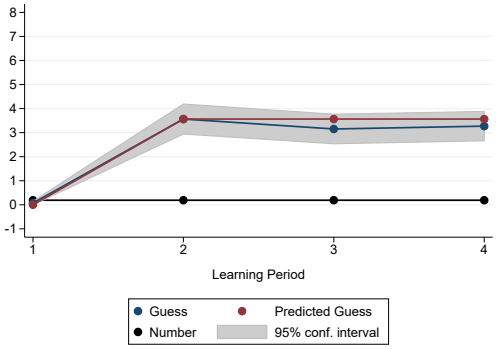
(C) Unbiased in MF rounds.



(D) Unbiased in SF rounds.



(E) Underconfident in MF rounds.



(F) Underconfident in SF rounds.

TABLE C.2. How well the model predicts the 3rd and 4th guess.

	All Rounds	Early Rounds	Late Rounds
Model	0.831***	0.826***	0.838***
	(0.025)	(0.030)	(0.030)
Const.	0.242**	0.238*	0.247*
	(0.091)	(0.119)	(0.096)
R^2	0.735	0.742	0.728
N	2052	1026	1026

	All Rounds		Early Rounds		Late Rounds	
	SF	MF	SF	MF	SF	MF
Model	0.834***	0.832***	0.818***	0.835***	0.854***	0.827***
	(0.027)	(0.033)	(0.040)	(0.040)	(0.030)	(0.045)
Const.	0.181	0.305**	0.260	0.238	0.104	0.362**
	(0.126)	(0.110)	(0.189)	(0.188)	(0.151)	(0.117)
R^2	0.758	0.697	0.751	0.706	0.767	0.684
N	1026	1026	542	484	484	542

	Overconfident	Unbiased Agents	Underconfident
Model	0.753***	0.890***	0.860***
	(0.046)	(0.056)	(0.037)
Const.	-0.261	0.554	0.282
	(0.179)	(0.286)	(0.149)
R^2	0.534	0.743	0.744
N	948	156	948

Classification of confidence types was based on elicited beliefs.

The dependent variable denotes subjects' actual guesses (the 3rd and 4th guess). The independent variable "Model" denotes guesses predicted by the model.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE C.3. The effect of revealed bias on learning in MF rounds.

	Classification of types based on revealed beliefs:			
	Overconfident or Unbiased		Underconfident or Unbiased	
	(1)		(2)	
Dependent variable: the difference between a guess and the number in MF rounds. Independent variables: dummy variables for each guess and their interactions.				
2^{nd} guess MF	-0.060	(0.260)	1.230^{***}	(0.338)
3^{rd} guess MF	-0.248	(0.327)	0.646^{**}	(0.300)
4^{th} guess MF	-0.487^{*}	(0.288)	0.455	(0.300)
Bias	-2.241^{*}	(1.150)	-15.419^{***}	(1.897)
Bias × 2^{nd} guess MF	-20.358^{***}	(0.976)	-16.713^{***}	(3.449)
Bias × 3^{rd} guess MF	-18.206^{***}	(2.480)	-3.182	(2.828)
Bias × 4^{th} guess MF	-19.542^{***}	(1.858)	-6.434^{**}	(2.448)
Const.	-0.038	(0.263)	-0.595^{**}	(0.285)
N	1348		1220	

"Bias" is based on beliefs revealed at the beginning of each round. It takes values between −1 and 1; positive values for overconfident and negative values for underconfident agents.

Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

APPENDIX D

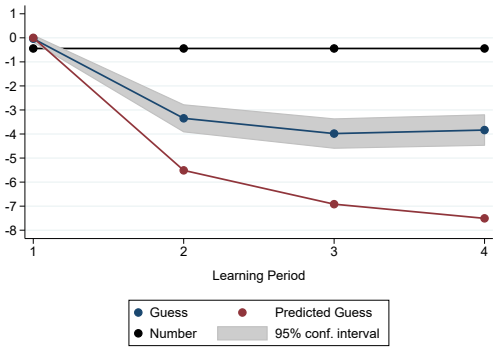
Ego-neutral Condition

In this section, we present the results from the ego-neutral control condition. First of all, we analyze the data in the same way as our main dataset: we re-do the analysis described in Section 1.3.2 in the paper using the data from the ego-neutral condition. Secondly, we combine the two datasets and analyze them jointly, complementing the results presented in Section 1.4 in the paper.

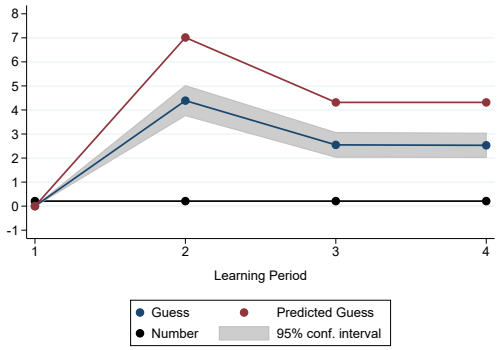
D.1. Misguided learning in the ego-neutral condition

In Figures D.1 and D.1, we present the learning outcomes of overconfident and underconfident participants in the ego-neutral condition. Tables D.1 and D.2 contain the results of the corresponding regressions, and in Table D.3 we gather the results of comparing pairs of coefficients. Overall, one can notice learning trajectories similar to those of overconfident and underconfident participants in the ego-relevant condition. A slight improvement could be spotted in the last guess of overconfident subjects in the multiple-feedback rounds. Those subjects seem to correct their choices in the direction of the true state. However, the correction is not significant at any acceptable level. Misguided learning is not eliminated in the ego-neutral condition, pointing towards the role of biased beliefs as its main source.

FIGURE D.1. The learning process in the ego-neutral control (MF rounds).



(A) Overconfident participants in multiple-feedback rounds.



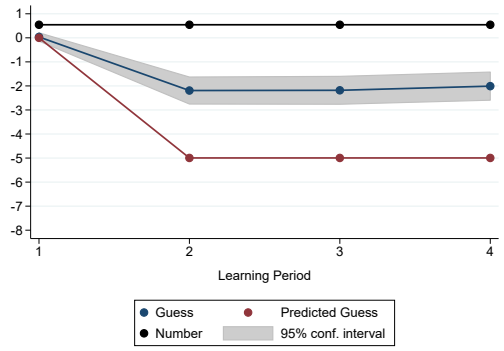
(B) Underconfident participants in multiple-feedback rounds.

TABLE D.1. The learning process in the ego-neutral control (MF rounds).

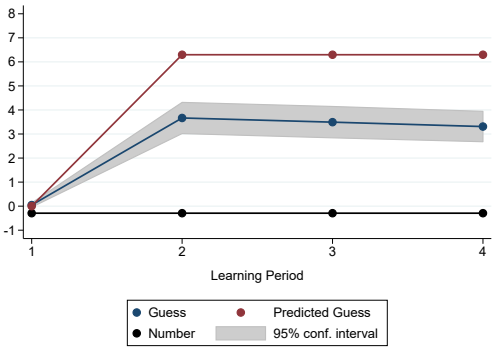
	Overconfident	Unbiased Agents	Underconfident
	(1)	(2)	(3)
Dependent variable: difference between a guess and the number in the SF rounds. Independent variables: dummy variables for each guess in the SF rounds.			
2^{nd} guess MF	-3.311***	0.407	4.388***
	(0.337)	(0.510)	(0.376)
3^{rd} guess MF	-3.945***	0.296	2.548***
	(0.404)	(0.629)	(0.290)
4^{th} guess MF	-3.799***	-0.074	2.530***
	(0.429)	(0.885)	(0.300)
Const.	0.406	-0.148	-0.210
	(0.756)	(0.749)	(0.246)
N	876	108	876

Note: The coefficients at the 2^{nd}, 3^{rd}, and 4^{th} guess SF remain unchanged if we control for subjects' relative performance (their actual position in the IQ test score distribution). Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

FIGURE D.2. The learning process in the ego-neutral control (SF rounds).



(A) Overconfident subjects in single-feedback rounds.



(B) Underconfident subjects in single-feedback rounds.

TABLE D.2. The learning process in the ego-neutral control (SF rounds).

	Overconfident (1)	Unbiased Agents (2)	Underconfident (3)
Dependent variable: difference between a guess and the number in MF rounds. Independent variables: dummy variables for each guess in the MF rounds.			
2^{nd} guess SF	-2.228***	1.185***	3.621***
	(0.341)	(0.324)	(0.352)
3^{rd} guess SF	-2.219***	1.000***	3.447***
	(0.353)	(0.225)	(0.358)
4^{th} guess SF	-2.045***	1.370***	3.265***
	(0.387)	(0.371)	(0.343)
Const.	-0.507**	-1.185**	0.337
	(0.222)	(0.494)	(0.250)
N	876	108	876

Note: The coefficients at the 2^{nd}, 3^{rd}, and 4^{th} guess MF remain unchanged if we control for subjects' relative performance (their actual position in the IQ test score distribution). Standard errors clustered at individual level. Their values in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.3. Comparison of the regression coefficients in the multiple- and single-feedback rounds in the ego-neutral condition.

(a) Overconfident Agents

	H_0: $\beta^2_{MF} \leq \beta^3_{MF}$	H_0: $\beta^3_{MF} \leq \beta^4_{MF}$	H_0: $\beta^2_{MF} \leq \beta^4_{MF}$
p-value	0.002^{***}	0.725	0.027^{**}
	H_0: $\beta^2_{SF} \leq \beta^3_{SF}$	H_0: $\beta^3_{SF} \leq \beta^4_{SF}$	H_0: $\beta^2_{SF} \leq \beta^4_{SF}$
p-value	0.524	0.890	0.879

(b) Unbiased Agents

	H_0: $\beta^2_{MF} = \beta^3_{MF}$	H_0: $\beta^3_{MF} = \beta^4_{MF}$	H_0: $\beta^2_{MF} = \beta^4_{MF}$
p-value	0.846	0.399	0.617
	H_0: $\beta^2_{SF} = \beta^3_{SF}$	H_0: $\beta^3_{SF} = \beta^4_{SF}$	H_0: $\beta^2_{SF} = \beta^4_{SF}$
p-value	0.282	0.174	0.184

(c) Underconfident Agents

	H_0: $\beta^2_{MF} \leq \beta^3_{MF}$	H_0: $\beta^3_{MF} = \beta^4_{MF}$
p-value	0.000^{***}	0.897
	H_0: $\beta^2_{SF} \leq \beta^3_{SF}$	H_0: $\beta^3_{SF} = \beta^4_{SF}$
p-value	0.070^{*}	0.009^{***}

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

D.2. Differences between subjects in the two conditions

As we have already mentioned in the paper, there is little difference between the treatment and the control group (see Table 1.6 in the paper) in the average relative performance or initial bias about own performance. If we look separately at the group of overconfident and underconfident participants (defined with respect to own performance) in the two conditions, there is a small difference in the performance of underconfident agents that is significant at the 10% level. Also, there are small differences in the initial bias within each group. We suspect that these differences are a consequence of having a relatively small sample. The exact values and tests in the two groups are gathered in Table D.4.

However, there are significant differences between overconfident subjects (with respect to own performance) in the ego-relevant condition and overconfident subjects (with respect to the other's performance) in the ego-neutral condition. One consequence of the random assignment of partners in the ego-neutral control is that the negative correlation between the decision-maker's performance and his bias is absent in this condition. The high (low) performing participants in the ego-neutral condition are not necessarily underconfident (overconfident) about the other's performance. As a result, the average performance of overconfident subjects in the ego-neutral condition is higher than that of the overconfident subjects in the ego-relevant condition, and the average performance of underconfident subjects in the ego-neutral condition is lower than that of the underconfident subjects in the ego-relevant condition. The exact values and tests for overconfident and underconfident agents could be found in TableD.5. We address this concern in the analysis by controlling for the performance of the decision-maker and his initial bias.

TABLE D.4. Differences between biased participants in the two conditions.

	Underconfident		**p-value**		
	Ego-neutral	**Ego-relevant**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Performance	0.817	0.775	0.935	0.130	0.065
	(0.021)	(0.018)			
Initial Bias	-0.233	-0.202	0.088	0.177	0.912
	(0.019)	(0.014)			
N	69	79			
	Overconfident		**p-value**		
	Ego-neutral	**Ego-relevant**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Performance	0.349	0.319	0.825	0.349	0.175
	(0.023)	(0.022)			
Initial Bias	0.256	0.293	0.090	0.181	0.910
	(0.019)	(0.020)			
N	71	79			

TABLE D.5. Differences between biased participants in the two conditions.

	Underconfident		**p-value**		
	Ego-neutral	**Ego-relevant**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Performance	0.634	0.775	0.000	0.000	1.000
	(0.032)	(0.018)			
Initial Bias	-0.218	-0.202	0.212	0.424	0.788
	(0.014)	(0.014)			
N	73	79			
	Overconfident		**p-value**		
	Ego-neutral	**Ego-relevant**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Performance	0.532	0.319	1.000	0.000	0.000
	(0.035)	(0.022)			
Initial Bias	0.247	0.293	0.043	0.086	0.957
	(0.018)	(0.020)			
N	73	79			

D.3. Learning in the ego-relevant and ego-neutral conditions

In this section, we present results complementing Tables 1.7 and 1.8 in the paper. In Tables D.6 and D.7, we present the regressions from Tables 1.7 and 1.8 in the paper controlling for the model's predictions (decisions implied by the model). The effect remains strong and significant for both overconfident and underconfident agents, with the regression coefficients similar to those in our initial specifications. Furthermore, we present the effect of the ego-relevant condition on learning in the remaining guesses – those not included in Tables 1.7 and 1.8 in the paper. In Tables D.8 and D.9, we show the results for the 2^{nd} and 3^{rd} guess of overconfident agents. The coefficients at the "Ego-relevant" variable in the 2^{nd} and 3^{rd} guess are slightly lower than the corresponding coefficients in the last guess (Table 1.7 in the paper) but remain positive and highly significant. In Tables D.10 and D.11, we present the results for the 3^{rd} and 4^{th} guess of underconfident agents. The difference between the ego-relevant and ego-neutral conditions in the 3^{rd} and 4^{th} guess is smaller than in the 2^{nd} guess. This should not come as a surprise: learning of underconfident agents is characterized by overshooting in the second guess, and one would expect the largest differences in decisions after the first feedback. Still, the sign of the effect in the 3^{rd} and 4^{th} guess is in line with our interpretation that underconfident agents become less mistaken about the state in the ego-relevant condition.

TABLE D.6. The effect of ego on learning of overconfident subjects.

	Dependent variable: the absolute difference between the 4^{th} guess and the number.		
	(1)	(2)	(3)
Ego-relevant	1.085**	1.632***	1.553***
	(0.520)	(0.510)	(0.243)
Controls 1	No	Yes	Yes
Controls 2	No	No	Yes
Controls 3	Yes	Yes	Yes
Observations	456	456	456

Note: The dependent variable is the absolute difference between the 4^{th} guess and the number. The sample includes only overconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. Controls 3 include the decisions implied by the model.

Standard errors clustered at the individual level. Their values in parentheses.

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

TABLE D.7. The effect of ego on learning of underconfident subjects.

	Dependent variable: the absolute difference between the 2^{nd} guess and the number.		
	(1)	(2)	(3)
Ego-relevant	-0.695*	-0.916**	-0.900***
	(0.396)	(0.385)	(0.371)
Controls 1	No	Yes	Yes
Controls 2	No	No	Yes
Controls 3	Yes	Yes	Yes
Observations	456	456	456

Note: The dependent variable is the absolute difference between the 2^{nd} guess and the number. The sample includes only underconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. Controls 3 include the decisions implied by the model.

Standard errors clustered at the individual level. Their values in parentheses.

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

TABLE D.8. The effect of treatment on learning (overconfident, 2nd guess).

	Dependent variable: the absolute difference between the 2^{nd} guess and the number.				
	(1)	(2)	(3)	(4)	(5)
Ego-relevant	0.969**	1.322***	1.294***	1.232***	1.019***
	(0.404)	(0.354)	(0.309)	(0.286)	(0.280)
Controls 1	No	Yes	Yes		
Controls 2	No	No	Yes		
Adjustment Type	Regression	Regression	Regression	Matching	Matching
Observations	456	456	456	456	456

Note: The dependent variable is the absolute difference between the 2nd guess and the number. The sample includes only overconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. In the matching estimator, observations are matched to the nearest neighbor based on the relative performance (Specification 4), and the initial bias and relative performance (Specification 5). In Specification 1-3, standard errors clustered at the individual level. In Specification 4-5, consistent standard errors as in Abadie and Imbens (2006). Their values in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.9. The effect of treatment on learning (overconfident, 3rd guess).

	Dependent variable: the absolute difference between the 3^{rd} guess and the number.				
	(1)	(2)	(3)	(4)	(5)
Ego-relevant	0.869*	1.330***	1.363***	1.228***	1.160***
	(0.520)	(0.495)	(0.443)	(0.367)	(0.347)
Controls 1	No	Yes	Yes		
Controls 2	No	No	Yes		
Adjustment Type	Regression	Regression	Regression	Matching	Matching
Observations	456	456	456	456	456

Note: The dependent variable is the absolute difference between the 3rd guess and the number. The sample includes only overconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. In the matching estimator, observations are matched to the nearest neighbor based on the relative performance (Specification 4), and the initial bias and relative performance (Specification 5). In Specification 1-3, standard errors clustered at the individual level. In Specification 4-5, consistent standard errors as in Abadie and Imbens (2006). Their values in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.10. The effect of treatment on learning (underconfident, 3rd guess).

	Dependent variable: the absolute difference between the 3^{rd} guess and the number.				
	(1)	(2)	(3)	(4)	(5)
Ego-relevant	-0.480	-0.664**	-0.514*	-0.657***	-0.452*
	(0.303)	(0.303)	(0.286)	(0.220)	(0.242)
Controls 1	No	Yes	Yes		
Controls 2	No	No	Yes		
Adjustment Type	Regression	Regression	Regression	Matching	Matching
Observations	456	456	456	456	456

Note: The dependent variable is the absolute difference between the 3rd guess and the number. The sample includes only underconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. In the matching estimator, observations are matched to the nearest neighbor based on the relative performance (Specification 4), and the initial bias and relative performance (Specification 5). In Specification 1-3, standard errors clustered at the individual level. In Specification 4-5, consistent standard errors as in Abadie and Imbens (2006). Their values in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.11. The effect of treatment on learning (underconfident, 4th guess).

	Dependent variable: the absolute difference between the 4^{th} guess and the number.				
	(1)	(2)	(3)	(4)	(5)
Ego-relevant	-0.344	-0.525*	-0.398	-0.515**	-0.300
	(0.320)	(0.308)	(0.290)	(0.225)	(0.212)
Controls 1	No	Yes	Yes		
Controls 2	No	No	Yes		
Adjustment Type	Regression	Regression	Regression	Matching	Matching
Observations	456	456	456	456	456

Note: The dependent variable is the absolute difference between the 4th guess and the number. The sample includes only underconfident participants. "Ego-relevant" indicates assignment to the ego-relevant condition (learning about own ability). Controls 1 include the relative performance of the decision-maker. Controls 2 include the initial bias of the decision-maker. In the matching estimator, observations are matched to the nearest neighbor based on the relative performance (Specification 4), and the initial bias and relative performance (Specification 5). In Specification 1-3, standard errors clustered at the individual level. In Specification 4-5, consistent standard errors as in Abadie and Imbens (2006). Their values in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

CHAPTER 2

Belief-based Utility and Signal Interpretation

People tend to overestimate their abilities and chances of success, making costly mistakes as they hold on to their biased beliefs at the expense of accuracy. This tendency, commonly referred to as *overconfidence*, generates significant costs for both the individual and society. A long-standing question in behavioral economics is how it can persist in environments with frequent feedback. In this paper, I explore one possible explanation.[1] I consider an agent who does not know his ability and receives a signal that either reveals it or not. The agent forms beliefs about both his ability *and* informativeness of the signal. Importantly, he values his beliefs about his ability, so that any change in these beliefs directly affects his utility function (Brunnermeier and Parker, 2005; Kőszegi, 2006; Caplin and J. V. Leahy, 2019). I attempt to answer the following questions: Does the agent perceive a favorable signal to be more informative than an unfavorable one? Would he perceive the signal differently if the signal did not affect his utility function?

To this end, I designed a simple experiment in which participants learn about their performance in an IQ test. In a treatment condition, participants received a signal about their performance and reported their beliefs about the signal's informativeness. I incorporate several changes to the classical design that allow me to better capture asymmetry in response to "good" and "bad" news. Moreover, I introduce a new control condition, in which participants decide about hypothetical signal realizations. They faced *the same* decision as subjects in the treatment condition but *without* receiving an actual signal. The difference between reports in the treatment and the control condition reveals the extent of belief manipulation in response to favorable and unfavorable signals, and pins down a causal effect of signal valence on updating. Moreover, it informs us about the underlying mechanism by showing how a change in beliefs (triggered by a signal) and the ensuing belief-based utility affect signal interpretation.

[1] Other explanations that are similar to my work (as they consider motivated reasoning rather than cognitive processes) can be divided into three categories: information avoidance (see Golman et al., 2017, for a comprehensive literature review), selective recall (Chew et al., 2020; Zimmermann, 2020; Huffman et al., 2022), and asymmetric updating. The last point mentioned comes the closest to my work and I review it in detail in the following section.

The data from the treatment condition shows that subjects perceive favorable signals as more likely to be informative. The average difference in the reported probability after a "good" versus a "bad" signal amounts to 13 percentage points and is significant at the 1% level. The result holds after controlling for potential selection. Moreover, the comparison between the treatment and the control condition indicates that the perception of a signal is significantly altered after receiving it. In the treatment condition, participants reported a 10.6 percentage points higher (a 27.9% increase) probability of a favorable signal being entirely informative about their performance. There is no significant difference after unfavorable signals. The inference about the signal has a lasting effect on subjects' beliefs about their ability. We observe additional asymmetry in how participants translate their beliefs about the signal into beliefs about ability. As a result, although signals significantly shifted subjects' beliefs, they did it selectively, and the aggregate overconfidence level remained virtually unchanged.

My study provides the first clear evidence of a causal effect of belief-based utility on signal interpretation. While the research on updating beliefs about ego-relevant traits has a long tradition, establishing causality has always been challenging. One difficulty lies in introducing exogenous variation in "ego-relevance": the way signals affect belief-based utility. Ideally, we would like subjects to receive the same feedback, but the feedback would have no *valence* – it would not be "positive" or "negative" in the sense that it would not bring participants additional belief-based utility. But how to separate feedback from its valence? Previous work focused on comparing how people update their beliefs about some ego-relevant parameter (e.g., one's performance in an IQ test) and how they update beliefs about some ego-neutral parameter (e.g., performance of a robot).[2] However, this comparison involves not only learning about ego-relevant and ego-neutral parameters, but also updating subjective beliefs, possibly multiple priors, and updating objective probabilities given by the experimenter. Treatment manipulation affects more than one aspect of the study undermining causal inference.

In this paper, I propose a novel experiment in which both the treatment and the control condition are based on the same subjective beliefs over the same ego-relevant characteristic. However, I introduce exogenous variation in how signals affect subjects' beliefs and their

[2]See, for instance, Eil and Rao (2011), Coutts (2019), and Möbius et al. (2022). One exception is a study by Buser et al. (2018), which compares how participants update beliefs about their performance in various tasks that differ in how relevant they are to the subject's self-esteem. However, in their set-up, it is not possible to introduce exogenous variation in ego-relevance. Grossman and Owens (2012) propose a control condition in which participants learn about the test result of another subject. In this case, subjects update their subjective beliefs about an unknown, ego-neutral variable.

belief-based utility: in the control condition, a signal is not realized, hence it does not affect subjects' beliefs nor their belief-based utility. Thereby, I separate feedback from its valence without changing other decision-relevant aspects of the design.

The study was conducted in August 2020 in the BonnEconLab at the University of Bonn. In total, I collected data from 222 participants. The experiment consisted of several parts. First, participants were given an IQ test and incentivized to do their best. After the test, they were asked to report their beliefs about their relative performance. Using an incentive compatible mechanism, I elicited subjective beliefs about one's test score falling into the 1^{st}, 2^{nd}, ..., 10^{th} decile of the score distribution. I referred to the deciles as "ranks", with 1 denoting the highest and 10 denoting the lowest rank.

After the belief elicitation, we described the framework to the subjects as follows: "There are two boxes. Box 1 contains 10 balls with numbers 1 to 10 written on them (each number occurs exactly once). Box 2 contains 10 balls with the same number written on every one of them. That number is equal to your rank." For example, if a subject's rank is 4, Box 2 contains 10 balls with the number "4" written on them.

In the main task, one ball was randomly drawn from one of the boxes (either box could be selected with equal probability) and presented to the subject. After seeing the ball, the participant reported his beliefs about the event that the ball came from Box 2 (with his rank). The report was made by dividing 100 points between the two boxes. I incentivized truthful reporting with the Binarized Scoring Rule (Hossain and Okui, 2013). The method was explained to the participants and they were informed that their chances to win the highest reward were maximized when they divided their points in a way that corresponded to their true beliefs about the box. We explained in intuitive terms how one can arrive at a Bayesian update given one's prior beliefs about the rank.

The design described above differs from experiments on belief updating in several ways.[3] First, I shift the focus from beliefs updating to subjects' inferences about the signal. I argue that updating takes two steps: assessing the information bore into a signal and incorporating it into prior beliefs.[4] I aimed at disentangling the effect of signal valence on the first step

[3] A detailed comparison of the designs used in the literature can be found in Appendix D.

[4] This holds true even in experiments that give participants a signal that is accurate with a certain probability (e.g., 75%). Before forming a posterior belief, one needs to answer the question of whether the observed realization reveals the state or should be attributed to noise.

from the way agents are aggregating information.[5] For this reason, I restricted the number of signals that participants receive to one.

Second, I use a richer state and signal space compared to previous studies. To understand why it is important, imagine a participant who believes that he is in the 80th percentile of the IQ test score distribution. Receiving a coarser signal, e.g., a signal indicating that his score was above the median, would not influence his beliefs as it merely confirms what he already knows. If the signal was more precise, e.g., it revealed that his score was only in the 60th percentile, it would affect his beliefs and, according to my hypothesis, induce a stronger reaction.

Last but not least, I define signal valence with respect to subjects' expectations. Being among 40% best performers is hardly good news if you expect to be among the top 10%. I incorporate this idea by defining a "good" signal to be the one above or equal to the median of individual belief distribution, which I elicited before the main task.[6]

An ideal counterfactual to the treatment condition would include a subject who has the same prior belief distribution (or the same set of prior belief distributions if the agent had multiple priors) and observes the same signal, but the signal has no effect on his belief-based utility function. To come as close as possible to the ideal counterfactual, I designed a control condition, which I describe below.

In the control condition, subjects do not see a ball being drawn, but are asked to report their beliefs about signal informativeness ex-ante, for every possible signal realization. The procedure, known as the Strategy Method, is commonly used in experiments investigating strategic interactions in games (Brandts and Charness, 2009). To alleviate concerns about the non-comparability of the two treatments, I adopted procedures that specifically targeted the issues raised in the literature.[7] I argue that a participant in the control condition faces

[5]Thus, my study is also related to the literature on self-serving attribution bias. It has been extensively studied by psychologists (see Mezulis et al., 2004, for a meta-analysis of the existing studies) and, more recently, by economists (Van den Steen, 2004; Coutts et al., 2020; Hestermann and Le Yaouanq, 2021). None of the studies, however, consider the counterfactual discussed in my paper.

[6]As a robustness check, I use different definitions of a "good" signal relative to beliefs: considering only signals that are strictly better than the median belief or replacing median with the mean.

[7]One concern raised in the experimental game theory literature is that players may gain a better understanding of the game if they are induced to think about the best strategies from the perspective of other players. One can imagine that considering every possible signal in the control condition could influence subjects' beliefs. I address this issue by presenting participants in the treatment condition with the screenshots from the control condition and asking them to consider every possible draw before they proceed to the main task. Moreover, I hope to alleviate another concern, the problem of framing the answers in the strategy method with the order of options, by randomizing the order of the signals presented to the subjects in both conditions.

the same decision as a subject in the treatment condition but without the signal affecting his beliefs and belief-based utility.[8]

The results lend support to the hypothesis that asymmetry in updating is due to an instantaneous reaction to signals. While there is a 6 percentage point difference in the beliefs reported after "good" versus "bad" signals in the control condition, the additional effect of a "good" signal in the treatment condition is almost twice as large (10 pp). I show that the effect strongly depends on the subjects' expectations. It is no longer present if a subject assigned zero prior probability to the rank indicated by the signal. Moreover, asymmetric updating about the box is followed by asymmetric updating about the rank. In the last part of the study, we again elicited subjects' beliefs about their rank (the entire belief distribution). The data reveal that participants translate their beliefs about the signal into beliefs about the rank in a motivated way, with those who received "good" signals being more consistent in their final reports. In the end, even though more participants received signals that were below their median beliefs, the average posterior belief in the sample was not significantly different from the average (overconfident) prior.

Using subjects' responses in questionnaires, I provide additional evidence to support my interpretation of the results as being driven by changes in belief-based utility. In the treatment condition, those participants who report experiencing hopelessness (a negative anticipatory emotion) tend to deviate more from the Bayesian benchmark. The effect is counteracted by the habitual use of emotion regulation strategies. Subjects who reported using more emotion regulation in their daily life tend to deviate less from Bayesian updating, even if they admit to feeling more hopeless. While only suggestive, the evidence supports the view that the treatment effect is stemming from the visceral, emotion-based reaction to signals that are indicative of a belief-based utility.

My work is based on the theoretical literature on overconfidence and belief formation. That literature postulates that people derive utility not only from physical outcomes but also from their beliefs about the current or future state (Brunnermeier and Parker, 2005; Kőszegi, 2006; Caplin and J. V. Leahy, 2019). The individual can choose his beliefs but faces a trade-off

[8]It is reasonable to assume that only realized signals induce subjects to revise their beliefs and bring them additional belief-based utility. The gain in utility can be sustained (or, in the case of unfavorable signals, mitigated) by distorting one's beliefs about signal informativeness.

between their accuracy (necessary to take the optimal action) and their desirability (a consequence of the non-monetary value beliefs bring to the agent). The tension is resolved by the agent manipulating his beliefs to the extent that he is not losing too much from actions taken based on those beliefs. Several studies demonstrated that agents significantly deviate from Bayes' rule when forming beliefs about their own intelligence or beauty (Eil and Rao, 2011; Ertac, 2011; Grossman and Owens, 2012; Buser et al., 2018; Coutts, 2019; Schwardmann and Van der Weele, 2019; Möbius et al., 2022). The main conclusion emerging from this strand of literature is that belief formation over ego-relevant characteristics significantly differs from learning about ego-neutral variables. At the same time, the direction of the effect and its magnitude vary across studies. The idea presented in this paper is related to research on emotions and decision-making (Lerner et al., 2015). One conclusion from the psychological literature is that emotions may influence decisions via changes in the content of thought, and vice versa. A similar hypothesis has been tested in a recent study of Engelmann et al. (2019) who investigate the impact of anxiety on wishful thinking. Using data from a carefully designed experiment, they show a causal effect of anticipatory anxiety on belief formation. Although I cannot argue about the causal impact of anticipatory emotions in my experiment, the suggestive evidence is in line with their findings.

The paper is organized as follows. The next section outlines the experimental design. In Section 2.2, I describe the main results. Section 2.3 presents the data from the final belief elicitation, and Section 2.4 describes the additional evidence. Section 2.5 concludes.

2.1. Experimental Design

The experiment consisted of two parts and is outlined in Figure 2.1. In the first part, subjects completed an IQ test intended to assess their cognitive ability. The second part included the elicitation of prior and posterior beliefs and a stage in which subjects received signals (or considered every possible signal realization in the control condition). I describe the procedures in detail in the following subsections.

2.1.1. IQ Test. In the first part of the experiment, I evaluated the subjects' cognitive ability using an IQ test.[9] The test consisted of 29 standard logic questions and participants were

[9] I decided to use intelligence as a basis for the learning exercise for several reasons. First, it is known that intelligence correlates strongly with educational achievement, success in the labor market, and income. Because of that, I expect people to care deeply about their cognitive ability. Therefore, IQ measure seems to be a good candidate for a genuine ego-relevant parameter. Second, the literature provides evidence that people have biased

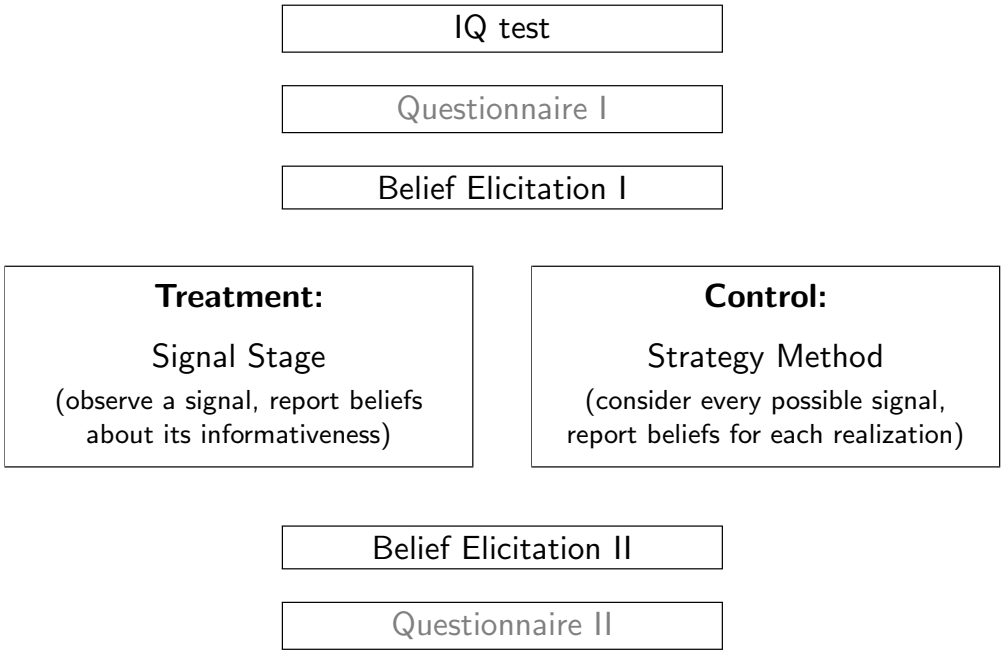


FIGURE 2.1. The outline of the experiment.

asked to solve as many of them as possible in 10 minutes. Individual scores were calculated based on the number of correctly answered questions minus the number of incorrect answers, and subjects were paid 0.75 Euro for every point they obtained.

Participants were informed that their earnings from the IQ test will be added to their earnings from the remaining parts of the experiment and paid at the end of the session. They were also informed that, although they will receive the entire sum of money at the end of the study, they will not learn immediately the exact number of points they obtained in the IQ test, nor how much money they earned in each part. Participants were informed that their IQ test results and the details of their payoffs will be available to them in one week after the session. Every participant received a personal link to a website on which his individual information was posted one week later.[10]

beliefs about their cognitive ability (with overconfidence prevailing among men), which suggests that learning about one's cognitive ability may be one of natural settings in which the mechanism is in play.

[10]This procedure served two purposes. First of all, I wanted to minimize dynamic concerns (e.g., subjects may adopt overly pessimistic beliefs to prepare themselves for the arrival of "bad news"). Second, this feature of the

2.1.2. Belief Elicitation. At the beginning of the second part, participants were told that they have to complete 3 tasks, for which they can earn up to 12 Euro. They were informed that *one task* will be drawn at random at the end of the session, and they will be paid only for that task.

In the first task, I elicited subjects' beliefs about their test scores being in the $1^{st}, 2^{nd}, ..., 9^{th}$, and 10^{th} deciles of the distribution of the test scores of 300 participants who took the same test in the BonnEconLab in previous sessions. I introduced 10 "ranks", with Rank 1 denoting the highest rank (assigned to participants whose IQ test scores were higher than or equal to the test scores of $90-100\%$ of all participants), and Rank 10 denoting the lowest rank (defined analogously). The first task was to allocate 100 points among the ranks in a way that reflects one's beliefs about the relative performance in the IQ test.

The screen-shot of the computer interface used by subjects is presented in Figure 2.2. Participants were allocating points by dragging blue arrows to selected positions. They were informed that they can move the arrows back and forth to correct their choices. The text below the scales informed a participant how many points are being allocated to a given rank and the allocation was immediately appearing on the graph to the right. The number above the graph indicated how many points the participant still has to allocate before he can proceed to the next task.

To incentivize truthful reports, I used the Binarized Scoring Rule following Hossain and Okui (2013). The random variable X can take one of 10 values: (1,0,...,0,0), (0,1,...,0,0), ..., (0,0,...,1,0), (0,0,...,0,1); the position of 1 indicates in which decile subject's IQ test score fell. After receiving agent's report $x = (x_1, ..., x_{10})$, where x_i denotes the share of points allocated to decile $i \in \{1, ..., 10\}$, I observed his IQ test score in the k^{th} decile, and the agent won the prize if the QSR for multiple events,

$$s(x,k) = 2x_k - \sum_i x_i^2 + 1,$$

exceeded a uniformly drawn random variable with the support $[0,2]$.

The formula was presented to the subjects in a simple way (avoiding mathematical notation). Importantly, I told participants the main implication of the method, that is, the probability of getting a large prize (12 Euro) is maximized when they allocate their points in a way that reflects their beliefs about their rank.

---

design enables me to collect data on who decided to check the test results. I describe the data on information acquisition in Appendix F.

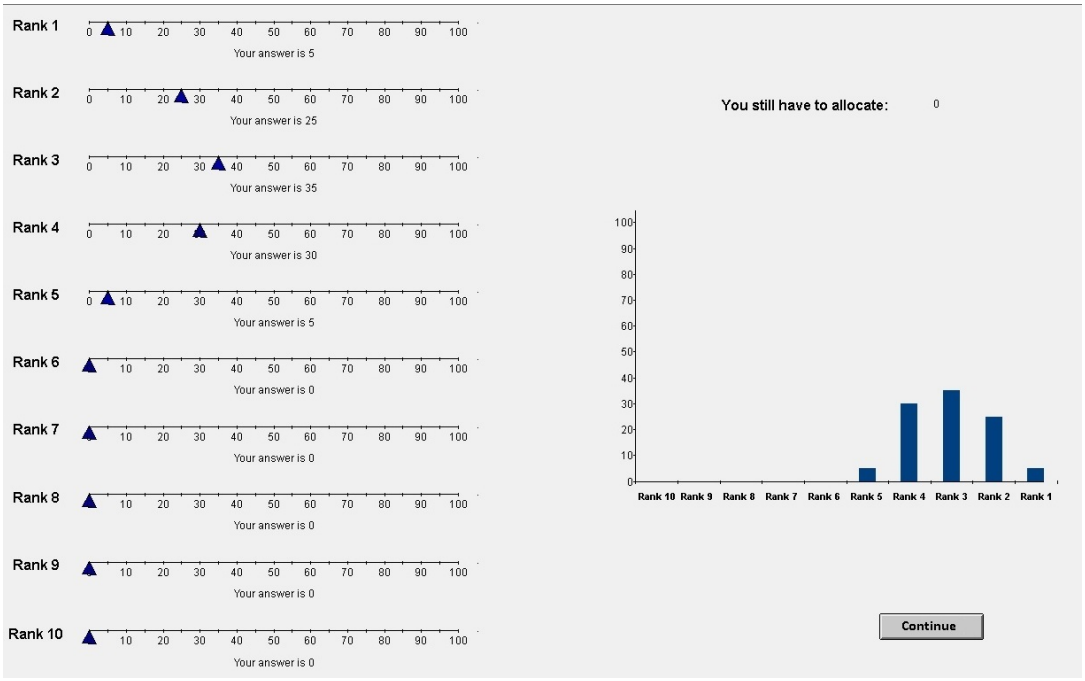


FIGURE 2.2. The screen-shot of the interface used in the first task.

I followed the same procedure during the second belief elicitation, after the signal stage (after the strategy method in the control condition). However, during the first belief elicitation, subjects were not aware that they will be asked to state their beliefs one more time.

2.1.3. The Signal Stage. After eliciting the prior beliefs, participants were given instructions for the second task. We explained the nature of the task in a simple language, using pictures and two illustrative examples. The task was framed in a neutral way and described as follows.

There are two boxes: Box 1 and Box 2. Each box contains 10 balls with numbers written on them. Box 1 contains balls with numbers from 1 to 10, and every number appears exactly once. The composition of the second box depends on the subject's rank in the IQ test. Box 2 contains 10 balls that all have one number written on them, and this number is equal to the individual rank. The composition of the boxes of a person assigned Rank 2 is presented in Figure 2.3.

For every participant, the computer program randomly selected one of the two boxes. Next, a ball was drawn from the selected box and displayed on the participant's screen. The participant did not know which box the ball was drawn from, but he knew that either box can

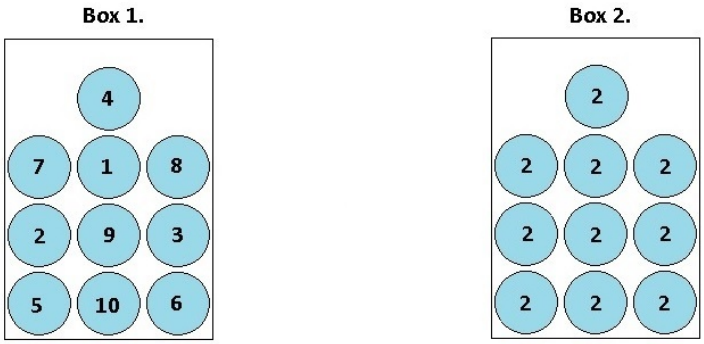


FIGURE 2.3. The composition of the boxes of a person whose rank was 2.

be selected with equal probability. After seeing the ball, he had to state his beliefs about the box selected by the computer.

I used the same incentive-compatible elicitation method as for the prior and posterior belief elicitations. Participants had 100 points to allocate between Box 1 and Box 2 in proportions that reflect their beliefs about the source of the signal, and were rewarded for the truthful report with a higher probability of getting a large prize (12 Euro).

Importantly, subjects were instructed how to arrive at the Bayesian posterior given one's prior belief distribution. I explained it with an example in two steps. First, I demonstrated how a person should allocate her points after different signal realizations if she knew precisely her rank. Then, I showed how a person should allocate her points if she was not sure about her rank, but was assigning a certain probability to it.

Step 1: How should a person ranked 2 allocate her points if she knew for sure that her rank is 2, and saw a ball with a number "2" on it? There are 10-times as many balls with "2" in Box 2 as there are in Box 1, hence it is 10-times as likely that the ball came from the second box. Therefore, the person should allocate 9 points to Box 1, and 10-times as many, 90 points, to Box 2 (the remaining point should be allocated to the box with higher probability).

Step 2: What if a person did not know her true rank, but she believed that there is 30% chance that her rank is 2? The same logic applies to this case. One can visualize 30% chance as 3 out of 10 balls in Box 2 having a number "2" on them.[11] In this imaginary case, there

[11] One reason why I decided to introduce 10 balls was the ease of exposition in a case when a person is uncertain about his rank.

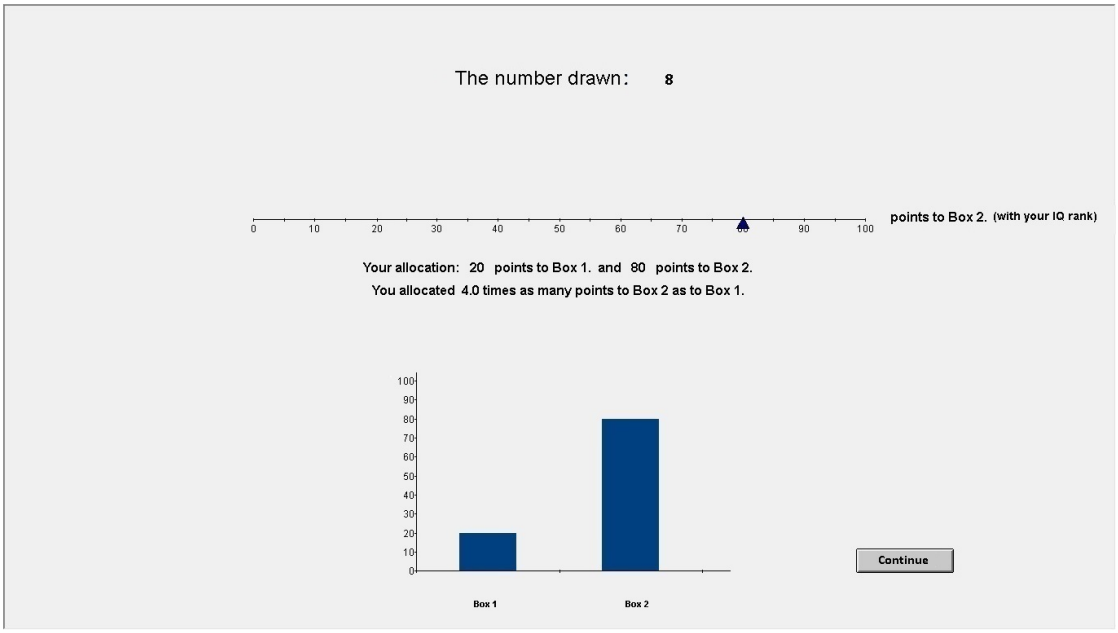


FIGURE 2.4. The screen-shot of the interface used in the second task.

are 3-times as many balls with the number "2" on them in Box 2 as in Box 1, implying an allocation of 25 points to Box 1 and 3-times as many (75 points) to Box 2.

The interface enabled subjects to split their points in desired proportions without calculating the respective ratios. The screen-shot of the interface used in the second task is presented in Figure 2.4. Crucially, the text below the scale informed subjects about their current allocation and the ratio between points allocated to the two boxes. By moving the cursor, participants could choose the number of points corresponding to allocating x-times as many points to one of the boxes (with $x \in \{1, 1.1, ..., 99\}$). The graph below was illustrating the current allocation.

Before proceeding to the signal stage, participants were required to answer a set of control questions, designed to check their understanding of the task (including the steps necessary for arriving at the Bayesian posterior). The control questions also pointed out the aspects that participants may have missed at the first reading, but were necessary to fully comprehend the task.

2.1.4. Experimental Conditions. I introduced two experimental conditions: treatment and control. In the control condition, subjects did not see the number that was drawn but

were asked to state their beliefs for every possible draw. The procedure, known as the Strategy Method, is commonly used in experiments investigating strategic interactions in games.

I informed participants in the control condition that the choices they are making are not entirely hypothetical. At the end of the session, one box was selected by the computer program and one ball was randomly drawn from the selected box. Subjects were paid as in the treatment condition, based on the decision that corresponded to the number drawn from the box. Note that the procedure is incentive-compatible as the probability of drawing any number is at least 5%.[12]

To alleviate concerns of the non-comparability of the two conditions, I adopted special procedures targeting the issues discussed in the literature. One concern raised in the experimental game theory literature is that players in the strategy method gain a better understanding of the game as a consequence of considering the problem from the point of view of different players. In my set-up, one can imagine that considering every possible signal realization may influence reported beliefs in the control condition.

For this reason, we asked the participants in the treatment condition to consider every possible signal realization *before* they saw the actual draw. Subjects were required to go through 10 slides, presented in random order, with the actual screen-shots of the interface displayed in the control condition. Participants were asked to contemplate a hypothetical decision in each slide before clicking on the button "Continue", which appeared on the screen only after 15 seconds. While only subjects in the control condition were allowed to enter their choices, both groups were required to go through the task.

Another problem that may arise in the Strategy Method is framing the answers with the order of options. I addressed the issue by randomizing the order of the numbers displayed to a subject in the control condition, and the order of slides presented to participants in the treatment.

2.1.5. Questionnaires. After each part of the experiment, I asked participants to fill in a 3-page questionnaire. The first set of questions, displayed on individual computer screens after the IQ test, included a short version of the Big-5 personality test (Gerlitz and Schupp, 2005) and the state-trait anxiety inventory STAI (Spielberger, 1983).

[12]However, if subjects were weighting the cost of cognitive effort against the expected payoff, they may exert less effort in the control condition. In this case, one would expect subjects to behave *less* rationally: their decisions would be characterized by a higher variance and they would end up further away from Bayesian update. This is the opposite of what I found.

The Big-5 personality test was designed to measure personality along five dimensions: extroversion, conscientiousness, openness to experience, neuroticism, and agreeableness. The STAI measures the current state of anxiety and anxiety level as a personal characteristic. The second set of questions, answered by the participants after the main task, comprised the Emotion Regulation Questionnaire (Gross and John, 2003) and a subset of questions from the Achievement Emotions Questionnaire (Pekrun et al., 2011).

The Emotion Regulation Questionnaire was designed to assess the habitual use of two strategies commonly used to alter emotions. To alleviate the emotional impact of a situation, one may try to reinterpret it in a different way. This emotion regulation strategy, broadly referred to as *reappraisal*, relies on "applying mental models to the often ambiguous and incomplete information" (Uusberg et al., 2019). The second emotion regulation strategy, *suppression*, involves "inhibiting ongoing emotion-expressive behavior" (Gross and John, 1998, cited in Uusberg et al., 2019).

People differ in their use of reappraisal and suppression, and these differences have implications for their experiences of emotions, behavior in response to those emotions, and general well-being (Gross and John, 2003). The habitual use of the two strategies is measured by the degree to which subjects agree with particular statements, e.g., "I keep my emotions to myself" or "When I want to feel less negative emotion, I change the way I'm thinking about the situation". I use the exact 10-item questionnaire developed by Gross and John (2003).

The Achievement Emotions Questionnaire was designed to measure *achievement emotions* (emotions that are directly linked to achievement activities or achievement outcomes) experienced by students in academic settings (Pekrun et al., 2011). I adopted part of the questionnaire to measure the following test-related emotions: enjoyment, hope, pride, relief, anger, anxiety, shame, and hopelessness.

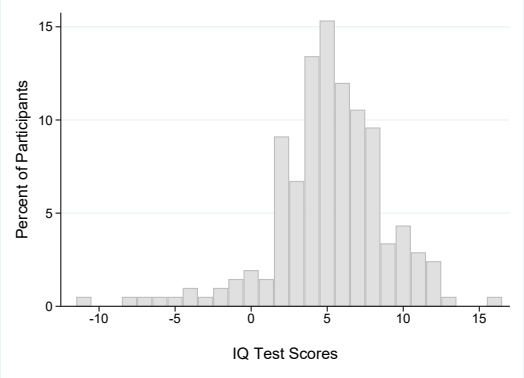
Participants in both conditions were asked to report what they felt *after* learning the nature of the task, but *before* they saw the number(s). They had to indicate, using a 7-point Likert scale, how strongly they agree (or disagree) with various statements, e.g., "I was proud of how well the test went", or "I was angry about the task I had to do".

2.2. Results

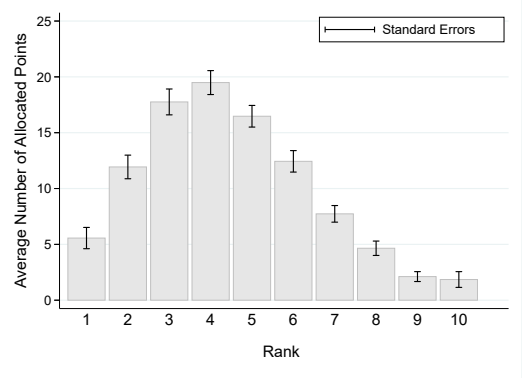
The experiment took place in August 2020 in the BonnEconLab at the University of Bonn.[13] I conducted 52 sessions, with 1 to 6 participants in each session. I collected data from 167 participants in the treatment condition and 55 participants in the control condition. The experiment lasted around 80 minutes and the participants earned 21.25 Euro on average. In the following section, I report the analysis based on the data from 209 participants who correctly answered at least half of the control questions (I excluded 13 participants, that is 5.8% of the sample).

2.2.1. IQ Test Results and Individual Ranks. Figure 2.5 presents the distribution of the IQ test scores and ranks assigned to the participants based on the test results. The IQ test score distribution is fairly symmetrical (skewness −0.83), with a mean of 5.13 and a standard deviation of 3.73. The average rank is 5.65 with a standard deviation of 2.67. Importantly, there is no significant difference in the average IQ test score or rank assigned to the participants in the treatment and control group (see Appendix A).

FIGURE 2.5. IQ Test Results and Beliefs.



(A) Distribution of test scores.



(B) Average number of points allocated to different ranks.

[13] Due to the Covid-19 pandemic, I followed special procedures to ensure the safety of participants and others involved. The number of participants per session was restricted to 6 to ensure each participant a place in a separate room. Desks, chairs, and computer equipment were disinfected after every session and the rooms were aired before every session for at least half an hour. At the time of the experiment (August 2020), the Covid-19 pandemic was mostly under control in Germany; the lockdown restrictions were eased, allowing restaurants, schools, and public places to open with appropriate safety measures.

2.2.2. Prior Beliefs about Rank. Before the main task, we elicited from every participant his entire belief distribution. I analyze the data in two ways. First, I look at the aggregate belief distribution. Then, I examine individual distributions and report the averages of individual measures (these include mean belief about rank, median and range). To look at the aggregate of individual belief distributions, I treat separately every decision to allocate x points, $x \in \{0,...,100\}$, to rank k, $k \in \{1,...,10\}$. For each of the 10 ranks, I calculate the average number of points allocated by the participants. The resulting aggregate distribution is presented on Panel (B) in Figure 2.5 (each bar indicates the average $+/-$ standard errors). It is visibly skewed to the right, with the mean belief of 4.47 and the median of 4. On average, the subjects appear to be *overconfident*, as they put a higher probability mass on lower (better) ranks.

In Table 2.1, I report the averages of individual measures of belief distribution. I look at the average mean belief, median belief, the first and third quartile, and range. Importantly, there is no significant difference between the treatment and the control group (see Appendix A). The averages, however, mask the fact that only 26 participants revealed symmetric belief distribution. Almost half of all subjects (100 participants) revealed a positively skewed belief distribution, and the remaining 83 participants revealed a negatively skewed belief distribution (the average difference between mean and median in both groups was 0.21). I define a person to be *overconfident* if his median belief is lower than his true rank. Similarly, I use a term *underconfident* to describe a person who assigns 50% or more probability mass to ranks higher than his true rank. A person is defined to be *unbiased* if his median belief matches his true rank.[14] Using this definition, there are 127 overconfident, 58 underconfident, and 24 unbiased participants in my sample. Importantly, there is no significant difference in the average bias (defined as a difference between the true rank and the median belief) between the treatment and the control group (see Appendix A).

TABLE 2.1. Individual belief distributions.

	Mean Belief	**Q1**	**Median**	**Q3**	**Range**
Mean	4.47	3.71	4.45	5.16	4.89
(Std. Dev.)	(1.75)	(1.74)	(1.79)	(1.87)	(1.57)

[14] In common language, Rank 1 denotes "the highest" rank, while Rank 10 is "the lowest". To avoid confusion, I will not use the customary phrases, but the terms that match the values (for example, a subject whose rank is 5 and median belief is 4 puts higher probability on *lower* ranks).

2.2.3. Decisions in the Main Task. The main experimental task, neutrally framed as "the second task", differed depending on the condition. In the treatment condition, subjects observed one number and reported their beliefs about the box from which the number was drawn. In the control condition, participants saw, in random order, numbers from 1 to 10, and stated a report for each one of them. In this section, I describe the raw data on subjects' decisions in the two conditions and present the results of the data analysis with and without using subjects' decisions in the control condition.

2.2.3.1. *Reports in the Treatment Condition (Raw Data).* First, I describe the raw data on the decisions made by participants in the second task. This was our main task: allocating points to Box 1 (with numbers from 1 to 10) and Box 2 (indicating one's rank) in a way that corresponds to one's beliefs about the source of the signal. I interpret points allocated to Box 2 as the probability that a subject assigns to the event that the number displayed on the computer screen is his rank.

Figure 2.6 presents the average number of points allocated to Box 2 after a signal received in the treatment condition. The numbers above the x-axis indicate how many participants received a given signal and stated a report. For example, 14 participants in the treatment condition saw "4" displayed on their computer screens and allocated, on average, 65 points to Box 2 (revealing the average subjective probability of 65% that the number "4" is their rank). It is useful to contrast these decisions with the Bayesian benchmark. For each participant, I calculated a Bayesian posterior about the box given his priors and signal realization.

FIGURE 2.6. Points allocated to Box 2 in the Treatment condition.

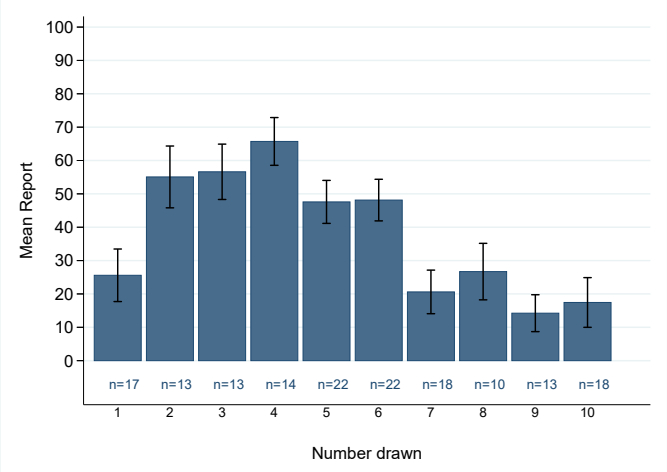
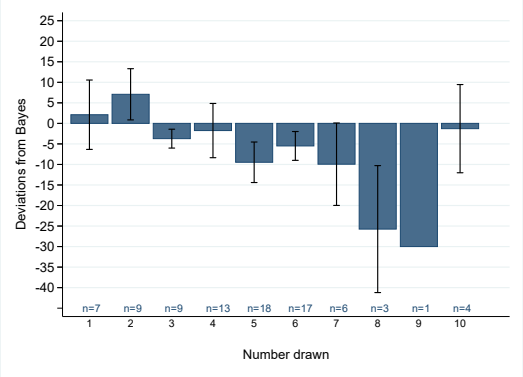
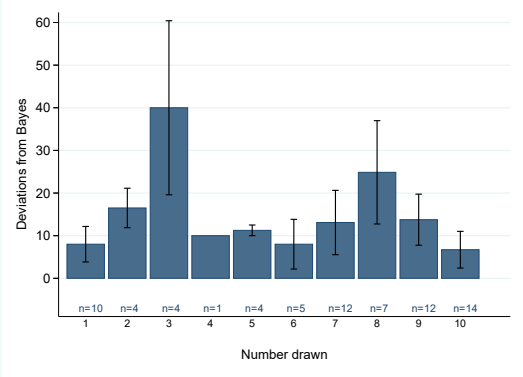


FIGURE 2.7. Mean deviation from Bayes for different signals in Treatment.



(A) Signals within priors.



(B) Signals outside priors.

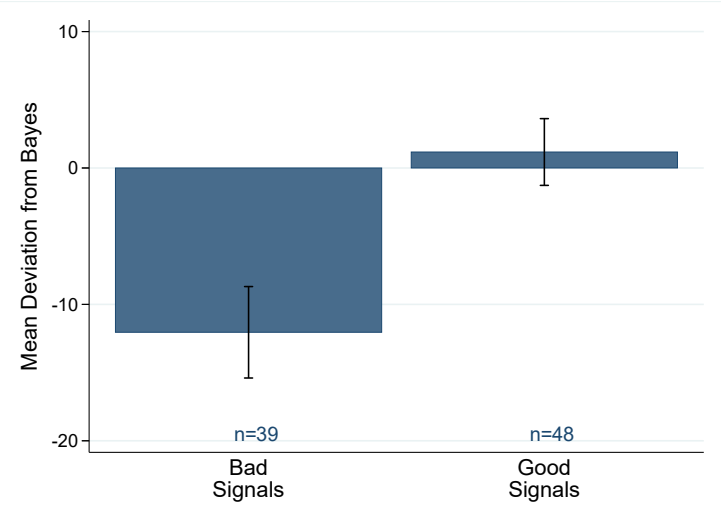
The average deviations from the Bayesian update in the treatment condition are presented in Figure 2.7. I separately plotted cases in which subjects assigned a non-zero prior probability to the number displayed on the screen (the graph on the left) and those in which subjects assigned the prior probability of zero (the graph on the right). I refer to the latter as "outside priors". One can notice that the average decisions are below zero for higher numbers – after worse signals, subjects tend to allocate fewer points than prescribed by the Bayes' rule. After better signals (indicated by lower numbers), subjects' decisions are closer to the bayesian benchmark.

However, by looking only at the signals' values one can miss an important point: signals might be perceived differently depending on subjects' expectations. A person who believes that her rank is "5" might perceive a signal "4" as a "good" signal. At the same time, a person who firmly believes that her rank is "1" can be disappointed after seeing a "4" and view it as a "bad" signal. We take this into account in the next section.

2.2.3.2. *Results Based on the Treatment Condition.* Our experimental design enables us to define the signal's valence depending on subjects' expectations.[15] In Figure 2.8, I present average deviations from the Bayesian update after signals that were worse than one's median belief (the left bar) and those that were better or equal to one's median belief (the bar on the right). Participants tend to allocate fewer points after signals that were worse than their median belief.

[15]Previous work on asymmetric updating mostly used binary state and signal space, and referred to the signal indicating a higher state as a "good" signal (see the literature review in Appendix D).

FIGURE 2.8. Deviations from Bayesian update after different signals.



The pattern visible on Figure 2.8 is confirmed by estimates presented in Table 2.2. The dependent variable is the number of points allocated to Box 2 (indicating one's rank). The independent variable "Bayes" denotes the number of points prescribed by the Bayes' rule. The variable "Good Signal" takes value 1 if the signal was lower or equal to the median belief and zero otherwise.[16] In the second column, we control for individual median belief, and in the last column, we add a control for individual rank (both variables could potentially influence the probability of receiving a "good" signal). The sample is restricted to the participants who assigned non-zero prior belief to the signals they received.[17] The coefficient at the "Good Signal" variable is around 13.0 and is significant at the 1% level, meaning that subjects report 13 percentage points higher beliefs that the signal is their rank after a "good" signal.

[16]The result is robust to using different definitions of a "good" signal relative to beliefs: considering only signals that are lower than the median belief as "good" signals, or replacing median with the mean of individual belief distribution.

[17]The estimation based on the entire sample and controlling for signals to which participants assigned zero prior probability ("outside priors") yield similar results, see Appendix B.

TABLE 2.2. The effect of the signal's valence.

	(1)	(2)	(3)
Bayes	0.956***	0.954***	0.961***
	(0.121)	(0.122)	(0.123)
Good Signal	13.629***	13.724***	12.976***
	(4.241)	(4.256)	(4.418)
Median Belief		-0.815	-0.537
		(1.167)	(1.245)
Rank			-0.604
			(0.914)
Constant	-9.527	-5.757	-3.819
	(7.651)	(9.383)	(9.862)
N	87	87	87

Standard errors in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 by participants in the treatment condition. "Bayes" denotes the number of points that should be allocated according to the Bayes' rule. The sample is restricted to subjects who received a signal to which they assigned non-zero probability. "Good Signal" indicator variable takes value 1 if the signal was below or equal to the median of subject's belief distribution, and 0 otherwise.

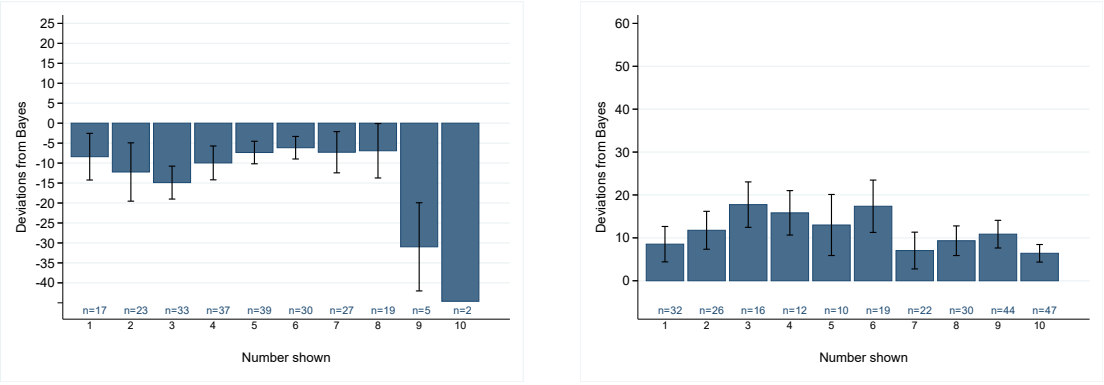
Importantly, this effect would not be captured if we defined "good" and "bad" signals in absolute terms. In the Appendix B.1, we replicate Figure 2.8 and Table 2.2 using the definition of "good" and "bad" signals commonly used in the literature: we define signals from 1 to 5 as "good" and signals from 6 to 10 as "bad". The effect is much lower and not significant at any acceptable level. The result points toward the importance of taking into account subjects' expectations to determine the signal's valence. Being among 50% best performers is hardly a good news if you expect to be among the top 10%. We later argue that asymmetric updating is mostly driven by an emotional reaction to signals, and one's prior beliefs likely serve as a reference point from which the signals are evaluated.[18]

[18] It is important to note that we gave participants much finer signals than most of the literature (usually signaling whether or not a subject is in the upper half of the distribution). We admit that subjects are likely to respond differently to coarser/finer signals and taking a simple average might not be a perfect comparison. Whether or not it is the case remains an open question for future research. Our hypothesis is that the average over finer signals is likely to be stronger than a response to a coarser signal due to a stronger emotional reaction: learning that one's performance is in the bottom 10% is likely to be more painful than a signal of being in the lower half.

2.2.3.3. *Data Analysis Using Control Condition.* In this section, I describe the data from the control condition. In Figure 2.9, I present participants' decisions separately for signals to which they assigned non-zero prior probability (Panel A), and those to which they assigned the prior probability of zero (Panel B). The averages are consistently below zero in the left panel, meaning that subjects tend to allocate fewer points than prescribed by the Bayes rule regardless of the signal under consideration. Note that the decisions were incentivized, thus allowing us to argue that participants made the best decisions using their prior beliefs about their rank and information from the signal. The only difference between the two conditions is that in the treatment condition participants received an actual signal.

In Table 2.3, I present the results of a regression analysis based on the data from both conditions. I restrict the sample to the participants who assigned a non-zero prior probability to the signal that appeared on their screen (the estimation based on the entire sample controlling for signals "outside priors" yielded similar results, see Appendix B). The dependent variable is the number of points allocated to Box 2. First, I regress it on the number of points prescribed by the Bayes' rule (the independent variable "Bayes") and a treatment dummy. As reported in the first column, both coefficients are positive and significant. In the second specification, I add an indicator variable "Good Signal", which takes value 1 if the signal was better or equal to one's median belief. A high and significant coefficient informs us that subjects tend to allocate more points to Box 2 in face of "good" signals. In the third specification, I add our main coefficient of interest – the interaction between the "Good Signal" and the "Treatment" variable. The coefficient at the interaction term is equal to 9.5 and

FIGURE 2.9. Mean deviation from Bayes for different signals in Control.



(A) Signals within priors.

(B) Signals outside priors.

TABLE 2.3. The effect of the signal's valence.

	(1)	(2)	(3)	(4)	(5)
Bayes	0.827***	0.765***	0.767***	0.767***	0.764***
	(0.093)	(0.094)	(0.093)	(0.093)	(0.092)
Treatment	5.761*	6.382**	1.012	1.015	1.005
	(2.982)	(2.884)	(4.111)	(4.125)	(4.170)
Good Signal		8.608***	5.944*	5.936*	5.943*
		(2.783)	(3.368)	(3.382)	(3.353)
Treatment × Good			9.474*	9.478*	10.247*
			(5.415)	(5.429)	(5.511)
Median Belief				0.048	-0.231
				(1.124)	(1.151)
Rank					0.661
					(0.596)
Constant	0.331	-1.126	0.381	0.164	-2.306
	(5.296)	(5.466)	(5.660)	(7.443)	(7.684)
N	319	319	319	319	319

Standard errors clustered at individual level. Their values in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 in the treatment condition. "Bayes" is the number of points that should be allocated according to the Bayes' rule. The sample is restricted to the participants who received (or considered) a signal to which they assigned non-zero probability."Treatment" is a variable indicating assignment to the treatment condition. "Good Signal" indicator variable takes value 1 if the signal was below or equal to the median of subject's belief distribution, and 0 otherwise.

significant at the 10% level. Importantly, the effect is similar if we add controls for individual rank and median belief.

One may worry about the fact that participants in the treatment condition with a probability of 50% decide about a signal that is their rank, while in the control condition, they decide about all 10 numbers. As a robustness check, we restrict the sample to the participants who saw a random number in the treatment condition. The results are gathered in Table B.4 in Appendix B. While the coefficient at the interaction term is not significant (p-value = 0.152), due to the small sample size, the coefficient of 9.8 is not different from the one presented in Table 2.3. We conclude that the differences in the probability of observing one's rank are not driving our results.

2.3. Belief Elicitation II

In this section, I take a closer look at beliefs about the rank elicited after the main task. I attempt to answer the following question: Do beliefs about the box translate to the posterior about the rank? Furthermore, I discuss the caveats of repeated belief elicitation and their consequences for the interpretation of the results.

2.3.1. Raw Data. Before delving into the analysis, I present the raw data on beliefs about the rank before and after the task. In Figure 2.10, I replicate Figure 2.5(B), juxtaposing the data from the first and the second belief elicitation. The graphs were created using only observations from the Treatment condition. There is little difference in aggregate beliefs before and after the signals. This result may seem surprising as it suggests that, on aggregate, subjects have not learned much, even though they received informative signals. I will show in the following section that it is not the case that our treatment manipulation failed to move participants' beliefs. Rather, it is a consequence of conservatism (updating too little in response to informative signals) and asymmetry (updating differently after negative and positive signals), as well as the fact that many of the "bad" signals that subjects received were outside their prior belief distributions.

First of all, I show that the signals indeed moved subjects' beliefs about the respective rank. In Figure 2.11, I plot the average number of points allocated to the rank indicated by

FIGURE 2.10. Average number of points allocated to 10 ranks.

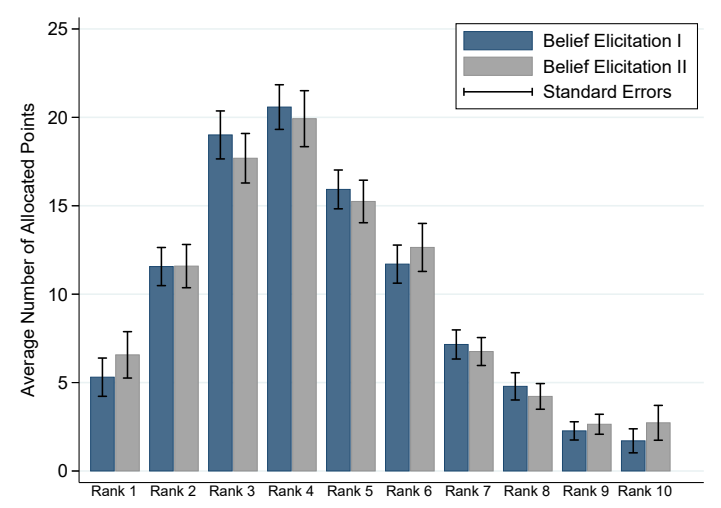
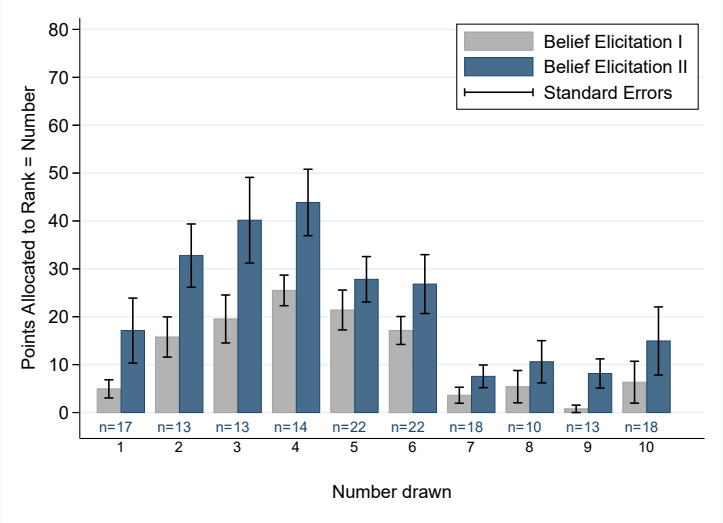


FIGURE 2.11. Points allocated to the relevant rank (before and after signals).



the signal (i.e., if a participant received a signal "2", only his allocations to Rank 2 are included). One can notice significant differences between the prior and posterior beliefs, and that those differences vary depending on the signal received. In Table 2.4, I present average allocations after "good" and "bad" signals separately for signals to which subjects assigned non-zero prior probability (I refer to them as "within prior") and those to which subjects assigned zero prior probability ("outside prior"). Two things are worth noting. First, although participants received more "bad" signals (n=91) than "good" signals (n=69), they received more "good" signals to which they assigned non-zero probability (n=48) than "bad" signals of a similar kind (n=39). Second, there is a larger change in beliefs after "good" signals within

TABLE 2.4. Points allocated to relevant rank ("good" and "bad" signals).

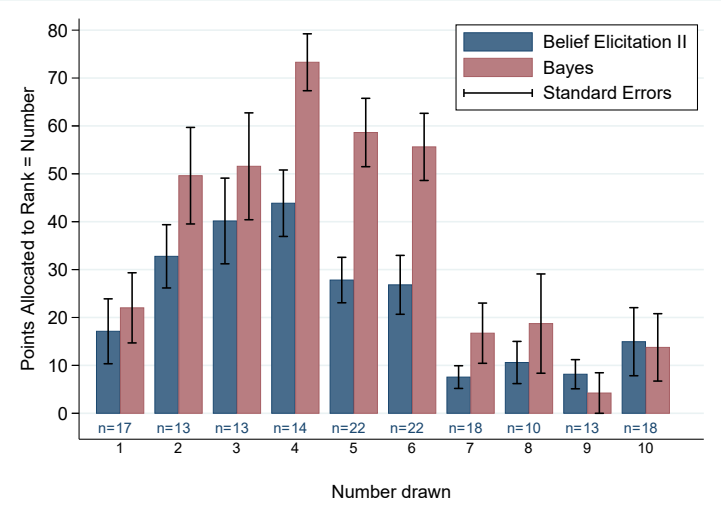
	"Good" Signals		**"Bad" Signals**	
	within prior	**outside prior**	**within prior**	**outside prior**
Belief Elicitation I	27.52	0	17.18	0
	(2.43)	(0)	(1.54)	(0)
Belief Elicitation II	47.75	6.52	23.77	6.08
	(4.14)	(2.73)	(2.89)	(1.62)
Difference	20.23	6.52	6.59	6.08
N	48	21	39	52

*Standard errors in parentheses.

priors than after "bad" signals within priors (20.23 versus 6.59 difference). In the following section, I analyze these differences in relation to the bayesian benchmark.

2.3.2. Data Analysis: Bayesian Benchmark. In Figure 2.12, I contrast beliefs elicited after the signal with the Bayesian benchmark, calculated based on the subject's prior beliefs about the rank.[19] If a subject assigned zero prior probability to the signal that he received I assume the benchmark to be zero. There are two things to be noted. First, subjects tend to allocate fewer points than prescribed by Bayes' rule. This could be a sign of conservatism, that is, under-reaction to new information. Second, the differences vary depending on signal realization. We use regression analysis to examine to what extent they are driven by differential responses to "good" and "bad" news.

FIGURE 2.12. Points allocated to the rank corresponding to the signal.



The estimation results are presented in Table 2.5. The dependent variable is the number of points allocated to the rank corresponding to the received signal. The first two columns report estimates based on observations from participants who received signals to which they assigned non-zero probability (signals "within prior"). The regression in the last column includes only participants who received signals to which they assigned a prior probability of zero ("outside prior"). In the first specification, I regress the dependent variable on the number of points they should have allocated according to Bayes' rule (the "Bayesian Posterior"

[19]Note that, since the signal is either entirely informative or uninformative, it should not affect any rank other than the one that corresponds to its realization. The prior beliefs on the relevant rank are all we need to calculate the Bayesian posterior.

TABLE 2.5. The effect of signal valence on beliefs about the respective rank.

	Signals "within prior"		**"outside prior"**
	(1)	(2)	(3)
Bayesian Posterior	0.921***	0.811***	
	(0.117)	(0.114)	
Good Signal		15.547***	0.447
		(4.361)	(3.088)
Constant	-26.735***	-27.687***	6.077
	(8.388)	(7.869)	(1.656)
N	87	87	73

Standard errors in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 by participants in the treatment condition. "Bayes" denotes the number of points that should be allocated according to the Bayes' rule. "Good Signal" indicator variable takes value 1 if the signal was below or equal to the median of subject's belief distribution, and 0 otherwise. The results are virtually unchanged if we control for individual rank and/or median belief.

variable). The coefficient at the "Bayesian Posterior" variable is 0.92 and statistically significant. Note, however, the negative coefficient at the constant variable, which informs us that participants allocated fewer points than they should have. In the second specification, we add the "Good Signal" variable, which takes value 1 if a signal was above or equal to one's median belief. The coefficient at the "Good Signal" variable is high and significant – participants tend to allocate 15.5 points more to the corresponding rank if they received a good signal. Thus, they revealed 15.5 percentage points higher beliefs that the signal is their rank after a "good" compared to a "bad" signal. The result remains the same if we control for individual rank and/or median belief (not shown in the table). There is no significant effect of signal valence for signals that were "outside" subject's prior belief distribution – the coefficient at the "Good Signal" variable is not significant in the last column in Table 2.5.

Several points should be kept in mind when interpreting the data from the second belief elicitation. First of all, one problem common in experiments measuring beliefs multiple times is that consistency motives may play a role. It has been shown in the literature (Falk and Zimmermann, 2017) that people prefer to act consistently in order to signal their skills to others. Despite our best efforts to ensure anonymity and instruct subjects to treat each part of the experiment independently, the second belief elicitation data may be tainted by

the desire to be seen as a consistent decision-maker.[20] If the consistency motives are in play, and people desire to make consistent reports in the two elicitation procedures, then what we found is a lower bound on the effect.

Second, while we explained to the subjects in intuitive terms how to arrive at a Bayesian posterior about the box, we provided no such guidance on how to translate the prior belief distribution and the signal to the posterior belief about the rank (nor we explained how to arrive at the posterior belief distribution given one's beliefs about the box). We believe this approach has both advantages and disadvantages. On the one hand, we did not frame participants in any way on what "should" be done in the experiment. On the other hand, we are losing control over what participants believe to be a rational course of action in an environment that is far from natural.[21] Yet, the posterior beliefs about the box and the belief distribution elicited in Belief Elicitation II are surprisingly consistent, lending credit to the use of these methods and corroborating the main results. We describe the comparison between the two in the following section.

2.3.3. Data Analysis: Consistency. During the experiment, participants' beliefs were elicited three times: before the task (Belief Elicitation I), as a part of the main task (beliefs about the box), and after the task (Belief Elicitation II). In Section 2.2, I described the data from Belief Elicitation I and subjects' beliefs about the box, while in the previous section, I contrasted Belief Elicitation I and II. There is one more comparison to be made, namely, beliefs about the box and Belief Elicitation II. This comparison enables us to answer the question: Do beliefs about the box translate to the posterior about the rank? The answer is important for the validity of our results – whether the asymmetry in beliefs about signal informativeness that we captured has any effect beyond the decisions in the main task.

To investigate this question, I construct a new variable "Consistent Posterior" that is a Bayesian posterior based on the subject's beliefs about the box. Then, I examine its relation to the posterior beliefs about rank. In Table 2.6, I present the results of a regression analysis based on observations from the treatment condition (those participants received actual signals), separately for signals to which subjects assigned non-zero prior probability (left side of the table) and signals to which subjects assigned zero prior probability (on the right). The

[20] This concern is alleviated in our main analysis, as it is based on a comparison between the Treatment and the Control, and there is no reason to believe that consistency motives differ in the two conditions.

[21] Although students are regularly given grades that are, to some extent, based on their relative performance, it is rather unusual to be asked to specify the entire belief distribution.

TABLE 2.6. The effect of signal valence on beliefs about the respective rank.

	"within prior"			**"outside prior"**		
	(1)	(2)	(3)	(1)	(2)	(3)
Consistent Posterior	0.783***	0.712***	0.537***	0.134**	0.134**	0.185**
	(0.076)	(0.082)	(0.117)	(0.063)	(0.063)	(0.075)
Good Signal		8.926**	-12.173		0.081	2.660
		(4.230)	(11.064)		(3.020)	(3.627)
Good × Consistent			0.332**			-0.175
			(0.161)			(0.138)
Constant	-14.142**	-14.436***	-5.045***	4.391***	4.368**	3.705*
	(5.354)	(5.250)	(6.883)	(1.874)	(1.808)	(5.354)
N	87	87	87	73	73	73

Standard errors in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 by participants in the treatment condition. The independent variable "Consistent Posterior" refers to the Bayesian prediction based on the belief about the box. "Good Signal" indicator variable takes value 1 if the signal was better or equal than the median of individual prior belief distribution, and 0 otherwise. The results are virtually unchanged if we control for individual rank and/or median belief.

dependent variable is the number of points allocated in Belief Elicitation II to the rank corresponding to the signal received.

The coefficient at the "Consistent Posterior" variable in Specification 1 tells us that this number is strongly related to the number of points subjects should have allocated given their beliefs about the boxes. While the coefficient is significantly lower than 1, it is clear that the beliefs about the box affect subjects' posterior beliefs about the rank. The relation is much weaker for the signals "outside" subjects' prior belief distributions. Moreover, for signals "within prior", there is a strong effect of a "good" signal, significant at the 5% level. Even after controlling for their decisions about the boxes, participants tend to allocate more points to the respective rank after a "good" signal, but only if they assigned a non-zero probability to the signal they received. In Specification 3, I add an interaction of the two variables. For signals "within prior", the coefficient at the interaction term is equal to 0.33 and significant at the 5% level. The results show that, in addition to motivated reasoning about the source of the signal, there is an asymmetry in translating those beliefs to the posterior beliefs about the rank, and participants who received "good" signals were more consistent in their final reports.

2.4. Additional Evidence

In this section, I examine a complementary data set of subjects' answers to questionnaires described in Section 2.1. First, I look at the subjects' personality traits, anxiety levels, as well as habitual use of emotion regulation strategies, and report their correlations with subjects' decisions in the second task.

2.4.1. Emotion Regulation Questionnaire. In Table 2.7, I report correlations between subjects' decisions in the treatment condition (relative to the Bayesian benchmark) and BIG-5 and STAI. The absolute deviations from Bayesian updating are correlated with the habitual use of reappraisal. The coefficient value of -0.18 indicates a weak, negative correlation significant at the 0.05 level. In Table 2.8, I present the estimates of regressions based on decisions made by participants in the treatment condition. I regress the independent variable, the absolute deviations from Bayesian update, on the independent variable "Reappraisal" that measures subject's habitual use of reappraisal. The coefficient at the "Reappraisal" variable is negative and significant at the 0.05 level. Reporting one point higher response on the 7-point Likert scale in questions about one's habitual use of reappraisal leads to a 3-point decrease in the distance from Bayesian update. The value doesn't change much if I control

TABLE 2.7. Correlations between the deviation from rationality and personal traits in the Treatment condition.

	DevB	Extr	Cons	Open	Neur	Agre	Trait	State	Reapp	Supr
DevB	1.00									
Extr	0.00	1.00								
Cons	0.05	-0.01	1.00							
Open	-0.09	0.22^*	0.10	1.00						
Neur	0.12	-0.24^*	-0.26^*	0.16^*	1.00					
Agre	-0.03	0.07	0.07	0.07	-0.13	1.00				
Trait	-0.07	0.29^*	0.35^*	-0.09	-0.71^*	0.23^*	1.00			
State	-0.15	0.28^*	0.17^*	-0.03	-0.58^*	0.24^*	0.70^*	1.00		
Reapp	-0.18^*	0.09	0.15	0.18^*	-0.17^*	0.22^*	0.13	0.17^*	1.00	
Supr	-0.04	-0.19^*	0.05	-0.17^*	-0.04	0.03	-0.13	-0.14	0.38^*	1.00

* $p < 0.05$

Note: "DevB" stands for deviations from Bayesian update. I use the labels: "Extr", "Cons", "Open", "Neur", and "Agre" for BIG-5 personality traits: extraversion, conscientiousness, openness to experience, agreeableness and neuroticism, respectively. I denote Anxiety trait and state with "Trait" and "State" (the two measures are defined such that a higher score indicates less anxious individual). "Reapp" and "Supr" stands for emotion regulation strategies: reappraisal and suppression.

for subject's rank, median belief or whether the signal he received was below or above his median belief or not within the prior belief distribution. The results show that subjects' decisions correlate with the way they handle positive and negative emotions in their daily life. The more used they are to regulate their emotions by thinking differently about the situation they found themselves in, the more they adhere to rational decision-making. To investigate this further, I take a closer look at emotion regulation strategies together with self-reported emotions experienced before the task.

TABLE 2.8. The effect of reappraisal on deviations from rationality.

	(1)	(2)
Reappraisal	-2.96**	-2.82**
	(1.29)	(1.29)
Constant	26.61***	27.33***
	(5.76)	(7.50)
Controls	No	Yes
Observations	160	160

Standard errors in parentheses

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

Note: The dependent variable is absolute deviations from the Bayesian update. Controls include the subject's rank and median prior belief, a dummy variable equal 1 if the signal was below or equal to the median prior belief, and a dummy variable equal 1 if the signal was outside of the subject's prior beliefs.

2.4.2. Test-related Emotions. In addition to the data presented so far, I collected survey data about test-related emotions experienced by participants before receiving the signal.[22]

Out of eight test-related emotions, anxiety and hopelessness significantly correlate with absolute deviations from Bayesian updating in the treatment condition. However, when I regressed absolute deviations from Bayesian updating on all test-related emotions, only hopelessness was highly statistically significant (p-value = 0.02) and remained so, even after adding additional controls on subjects' rank, median belief, and signal's value or its relation to the subject's beliefs.

[22] In the instructions displayed on the screen, I highlighted that questions refer to the particular moment in time: *after* learning the nature of the task, but *before* seeing the number.

Hopelessness was measured by agreement with the statement “I felt that I would rather not do this part because I’ve lost all hope.”. As reported in the first column in Table 2.8, stating a 1-point higher answer to the question translates to an increase of 4.3 points in absolute deviation from Bayesian updating (controlling for all remaining test-related emotions). The coefficient at the “Hopelessness" variable remains unchanged if I control for the emotion regulation strategies: suppression and reappraisal (Specification 2) in Table 2.8. Of the two strategies, only reappraisal is different from zero and significant. Moreover, it has the expected negative sign and value similar to that reported in Table 2.4. I hypothesize that the use of reappraisal counteracts the negative impact of Hopelessness. To test this hypothesis, I add to the regression the interaction of “Hopelessness" and “Reappraisal". I report the estimation results in the last column of Table 2.8. The coefficient at the interaction term is negative and highly significant, whereas the coefficient at the “Reappraisal" variable loses its significance. At the same time, the coefficient at Hopelessness increases fourfold and gains significance, suggesting that its impact is much larger without the offsetting effect of reappraisal.

TABLE 2.9. The effect of emotions on deviations from rationality.

	(1)	(2)	(3)
Hopelessness	4.31**	4.30**	17.23***
	(1.83)	(1.82)	(4.62)
Reappraisal		-2.82**	2.21
		(1.42)	(2.16)
Hopelessness × Reappraisal			-3.10***
			(1.02)
Constant	10.00	20.18**	2.73
	(8.28)	(10.11)	(11.40)
Controls 1	Yes	Yes	Yes
Controls 2	No	Yes	Yes
Observations	160	160	160

Standard errors in parentheses

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

Note: The dependent variable is absolute deviations from the Bayesian update. The independent variable “Hopelessness” was measured by the extent to which a subject agreed with the statement “I felt that I would rather not do this part because I’ve lost all hope.”. “Reappraisal” refers to self-reported habitual use of reappraisal. Controls 1 include all other emotions reported by subjects; Controls 2 include the measure of habitual use of suppression.

While only suggestive, the evidence presented in this section supports the view that the treatment effect is stemming from the visceral, emotion-based reaction to signals. That reaction lies at the heart of what economists call "the belief-based utility" and is the driving force behind asymmetric updating.

2.5. Conclusions

In this paper, I propose a new test of the hypothesis that people interpret favorable feedback as more informative. To this end, I designed a simple experiment with two conditions. In the treatment condition, participants observe a signal about their intelligence and decide whether the signal is informative or not. In the control condition, participants make the same choice without receiving a factual signal: they are asked to specify their actions conditioning on possible signal realizations. This design allows me not only to pin down the causal effect of signal valence on updating but also to uncover the underlying mechanism. The experimental data reveal that people tend to interpret favorable signals as more informative due to the changes in belief-based utility. Participants reported a 10 percentage point higher probability of a positive signal being entirely informative about their rank after receiving it, compared to what they would conclude ex-ante, without observing its realization. The results cast a new light on the origins of overconfidence, pointing towards the role of affect in asymmetric updating. Moreover, we observe additional asymmetry in how subjects translate their beliefs about signal informativeness into beliefs about ability – participants who received "good" signals were more consistent with their previous reports. Even though signals significantly shifted subjects' beliefs, they did it selectively, with "good" signals having a larger impact on final beliefs. As a result, the aggregate overconfidence level remained the same at the end of the experiment.

APPENDIX A

Differences between the treatment and the control group

TABLE A.1. Differences between participants in Treatment and Control.

			p-value		
	Treatment	**Control**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
IQ score	5.12	5.16	0.47	0.94	0.53
	(0.30)	(0.50)			
Rank	5.59	5.82	0.31	0.61	0.69
	(0.21)	(0.39)			
Bias	1.18	1.23	0.46	0.91	0.54
	(0.22)	(0.43)			
Absolute Bias	2.38	2.60	0.24	0.47	0.76
	(0.14)	(0.28)			
N	160	49			

Note: "Bias" is defined as difference between rank and median belief. Standard errors in parenthesis.

TABLE A.2. Differences in prior belief distributions (Treatment vs Control).

			p-value		
	Treatment	**Control**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Prior Beliefs:					
Mean	4.43	4.56	0.33	0.65	0.67
	(0.14)	(0.26)			
1^{st} Quartile	3.69	3.79	0.35	0.70	0.65
	(0.13)	(0.27)			
Median	4.41	4.58	0.28	0.56	0.72
	(0.13)	(0.27)			
3^{rd} Quartile	5.11	5.34	0.23	0.45	0.77
	(0.15)	(0.27)			
N	160	49			

TABLE A.3. Deviations from Bayes in the main task (Treatment vs Control).

	Dependent variable: absolute difference between subjects' reports and the Bayesian benchmark.
	(1)
Treatment	-0.46
	(1.72)
Constant	14.23***
	(0.93)
Observations	650

Standard errors clustered at the participant level.

Their values in parentheses.

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

Note: The dependent variable is the absolute difference between subjects' belief about the box and the Bayesian benchmark (in cases when subjects' assigned zero prior probability to the signal displayed on-screen the rational benchmark is assumed to be 0). I interpret the dependent variable as a measure of rationality demonstrated during the task. "Treatment" is an indicator variable taking value 1 if the subject was in the Treatment condition and 0 otherwise (the Control condition).

APPENDIX B

Additional Results

B.1. Defining signal valence in absolute terms

FIGURE B.1. The average deviation from Bayes for signals above/below 5.

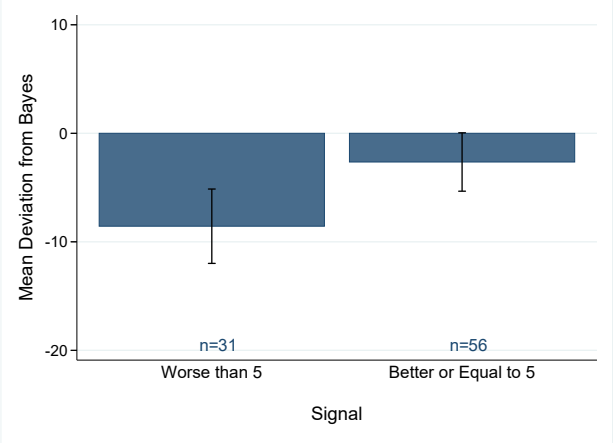


TABLE B.1. The effect of the signal's valence defined in absolute terms.

	(1)	(2)
Bayes	1.037***	1.048***
	(0.125)	(0.125)
Good Signal (below 5)	5.699	3.446
	(4.511)	(4.965)
Rank		-1.038
		(0.960)
Constant	-10.732	-4.456
	(8.189)	(10.030)
N	87	87

Standard errors in parentheses

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

B.2. Results based on the entire sample

In this section, we replicate the results from Table 2.2 and Table 2.3 using the data from the entire sample. The results show that the coefficients at the "Good Signal" variable and the interaction term are very similar to the ones reported in the main text.

TABLE B.2. The effect of the signal's valence in the Treatment condition.

	(1)	(2)	(3)	(4)	(5)
Bayes	0.713***	0.992***	0.956***	0.956***	0.967***
	(0.051)	(0.129)	(0.130)	(0.131)	(0.131)
Good Signal	9.694***	8.958**	13.629***	13.662***	12.320***
	(3.496)	(3.461)	(4.558)	(4.573)	(4.616)
Outside Prior		20.239**	22.315**	22.041**	23.072***
		(8.633)	(8.694)	(8.775)	(8.745)
Outside Prior × Good Signal			-10.893	-10.302	-9.745
			(6.961)	(7.289)	(7.255)
Median Belief				-0.283	0.260
				(1.004)	(1.050)
Rank					-1.079*
					(0.644)
Constant	9.231***	-9.241	-9.527	-8.217	-4.970
	(2.515)	(8.260)	(8.224)	(9.466)	(9.609)
N	160	160	160	160	160

Standard errors in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 by participants in the treatment condition. "Bayes" denotes the number of points that should be allocated according to the Bayes' rule (or zero if a subject assigned zero prior probability to the signal displayed on screen). "Good Signal" indicator variable takes value 1 if the signal was below or equal to the median of subject's belief distribution, and 0 otherwise. "Outside Prior" indicator variable takes value 1 if a subject assigned zero prior probability to the rank corresponding to the signal he received.

TABLE B.3. The effect of the signal's valence in the two conditions.

	(1)	(2)	(3)	(4)	(5)
Bayes	0.695***	0.670***	0.670***	0.767***	0.765***
	(0.039)	(0.039)	(0.039)	(0.093)	(0.093)
Treatment	4.547**	4.841**	1.616	1.012	1.015
	(1.952)	(1.914)	(2.601)	(4.105)	(4.136)
Good Signal		5.0767**	3.290	5.944*	5.924*
		(2.140)	(2.587)	(3.362)	(3.361)
Treatment × Good			7.357*	9.474*	9.875*
			(5.415)	(5.407)	(5.437)
Outside Prior				9.962*	10.004*
				(5.890)	(5.951)
Controls 1	No	No	No	Yes	Yes
Controls 2	No	No	No	No	Yes
Constant	9.494	7.853	8.685	0.381	-1.477
	(1.291)	(1.464)	(1.640)	(5.651)	(6.249)
N	650	650	650	650	650

Standard errors clustered at individual level. Their values in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 in the treatment condition. "Bayes" is the number of points that should be allocated according to the Bayes' rule. The sample is restricted to the participants who received (or considered) a signal to which they assigned non-zero probability."Treatment" is a variable indicating assignment to the treatment condition. "Good Signal" indicator variable takes value 1 if the signal was below or equal to the median of subject's belief distribution, and 0 otherwise. "Outside Prior" is an indicator variable taking value 1 if a subject assigned a probability of zero to the signal. Controls 1 include interactions of the "Outside Prior" variable with "Treatment" and "Good Signal". Controls 2 include individual rank and median belief.

B.3. Results based on a restricted sample

TABLE B.4. The effect of the signal's valence (restricted sample).

	(1)	(2)	(3)	(4)	(5)
Bayes	0.795***	0.735***	0.731***	0.732***	0.731***
	(0.101)	(0.104)	(0.104)	(0.104)	(0.103)
Treatment	8.551**	9.140**	3.668	3.480	3.173
	(3.669)	(3.508)	(5.379)	(5.392)	(5.458)
Good Signal		7.696**	6.323*	6.260*	6.226*
		(3.021)	(3.338)	(3.363)	(3.341)
Treatment × Good			9.799	10.005	10.019
			(6.784)	(6.839)	(6.952)
Median Belief				0.351	0.131
				(1.230)	(1.240)
Rank					0.629
					(0.642)
Constant	2.311	1.230	2.299	0.700	-2.023
	(5.850)	(5.996)	(8.102)	(7.443)	(8.526)
N	270	270	270	270	270

Standard errors clustered at individual level. Their values in parentheses.

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

Note: The dependent variable is the number of points allocated to Box 2 in the treatment condition. "Bayes" is the number of points that should be allocated according to the Bayes' rule. The sample is restricted to the participants who were not guessing their own rank. "Treatment" is a variable indicating assignment to the treatment condition. "Good Signal" indicator variable takes value 1 if the signal was below or equal to the median of subject's belief distribution, and 0 otherwise.

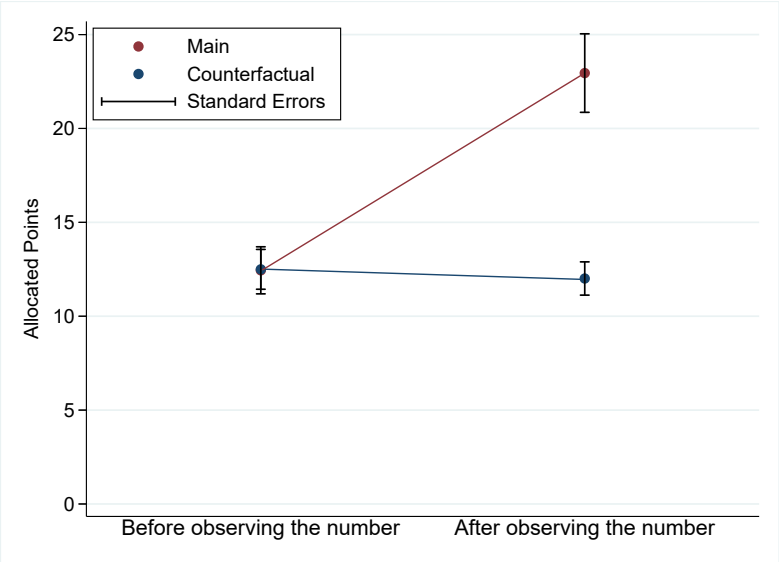
APPENDIX C

Manipulation Check

I argue that the treatment effect is caused by the utility from beliefs induced by the signal. First of all, I provide evidence that the signal received in the treatment condition affected subjects' beliefs. In Figure C.1, I present subjects' beliefs before and after the main task, that is, beliefs revealed in the first and the second belief elicitation. The graph shows points allocated to the rank corresponding to the number displayed on subjects' screens. I compare these values to the counterfactual: how many points they would allocate to the respective ranks if they did not receive a signal. There is no change in beliefs in the counterfactual scenario (denoted with a blue line). In the treatment condition, the change in beliefs is significant (marked in red on the graph). Subjects allocated almost two times as many points in the second belief elicitation to the rank displayed on the screen.

Second, I exclude alternative hypotheses. One may worry that subjects in the control condition exerted less effort per decision (e.g., due to increasing marginal cost of effort or

FIGURE C.1. Beliefs before and after the signal.



lower monetary incentives in the control condition). To alleviate this concern, we asked participants in the treatment condition, before they received an actual signal, to consider about every possible signal realization. We showed them, one by one, every possible number and asked them to think what they would do if this number was drawn later. This additional part makes the total time spent on the second task similar in both conditions. One may argue that the total time spent on the task may not be a perfect measure of effort and there still may be differences in cognitive effort exerted when making a decision in the treatment and in the control condition. However, if this was the case, one would expect larger deviations from the rational benchmark in the control condition. As reported in Table A.3 in Appendix A, there is no significant difference in absolute deviations from the Bayesian benchmark in the two conditions. I provide additional evidence to support my interpretation of the results as being driven by changes in belief-based utility in Section 2.4.

APPENDIX D

Literature: Design Comparison

The experiment developed for this paper differs from designs used in the literature in several ways. First of all, the new control condition addresses the problem of causal identification of the effect of signal valence, as described in the main body of the paper. Second, also the treatment condition diverges from the paradigm commonly used in experimental studies on belief formation. Guided by the hypothesis that it is the belief-based utility what drives the updating about ego-relevant characteristics, I aimed at designing an updating task that induces a strong emotional reaction to the signal. In order to clarify the differences between my design and experiments conducted in the past, I gathered and described dissimilar features of the design in Table D.1.

While there are many papers studying overconfidence and asymmetric updating, in this review, I focus on papers that study updating about ego-relevant characteristics and do so by asking subjects to update their beliefs about their *relative* performance. For a review of the beliefs updating literature that includes updating about absolute performance as well as updating about non-ego-relevant parameters, I refer the reader to the recent works of Barron (2021) and Coutts (2019). An even broader review of the literature on errors in probabilistic reasoning could be found in Benjamin (2019).

The papers gathered in the first column in Table D.1 are categorized based on various design features. In the second column, I describe the corresponding design feature used in my experiment. The last column presents the rationale behind choosing this particular feature for my work. One important design feature that requires an additional comment is the information structure. In almost all of the work reviewed in this section, the information structure follows the scheme presented in Figure D.1.[1] There are two states of the world H and L indicating whether one's score was in the upper or the lower half of the test score distribution, and each subject receives a signal that is informative about the state with known precision,

[1] See Table D.1 for the references. Two papers that deviate from this signal structure are Eil and Rao (2011) and Zimmermann (2020) who introduce 10 states of the world and binary signals. A signal informs a subject whether or not he ranked higher than another participant who was randomly drawn from a group of 10 (see Figure D.3; I denote the signals with H and L). The signal precision depends on the state and, for the first signal, can take one of the values: 55.6%, 66.7%, 77.8%, 88.9% or 100% (for the second signal it is 50%, 62.5%, 75%, 87.5% or 100%, as comparisons are made without replacement).

e.g., 75%, as shown in Figure D.1. However, this signal structure becomes more complicated if extended to a larger signal and state space (see Figure D.2) and I am not aware of any experimental work that implements it. The papers that used 10 states of the world in their design, Eil and Rao (2011) and Zimmermann (2020), use binary signals (see Figure D.3).

FIGURE D.1. Design used in the literature (2 states).

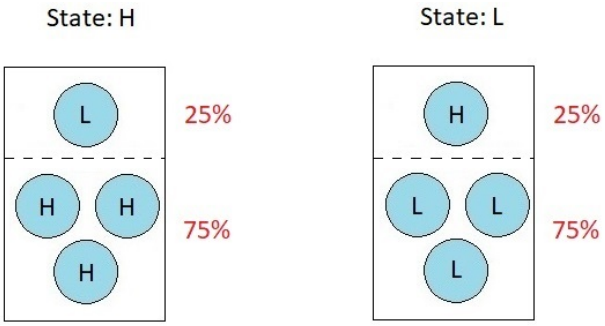


FIGURE D.2. Design used in the literature extended to 10 states.

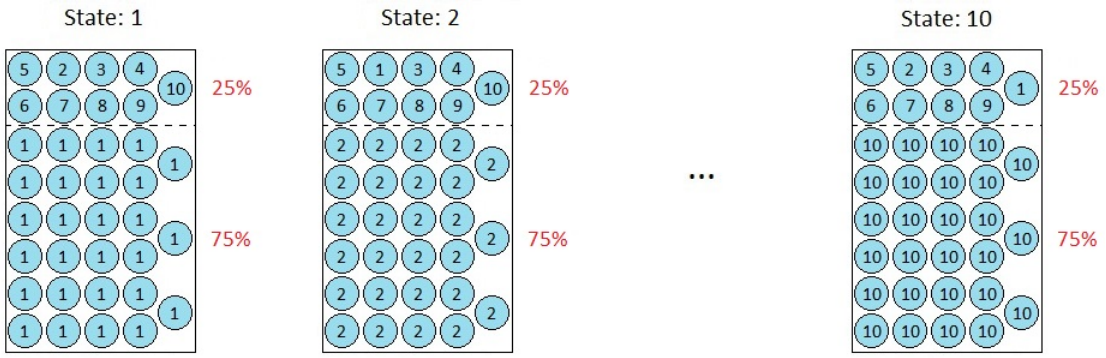
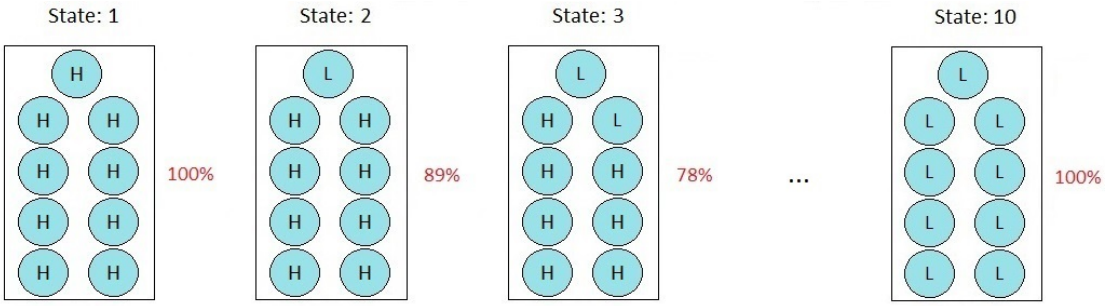


FIGURE D.3. Design used in Eil and Rao (2011) and Zimmermann (2020).



The design used in the literature extended to 10 states (Figure D.2) can be simplified by distinguishing two urns: one with balls indicating the state ("IQ" urn), and the other with every possible number ("Random" urn).[2] This is presented on Figure D.4 and Figure D.5 that illustrate the cases of 2 and 10 states of the world, respectively. Note that the information structure introduced in Figure D.4 *is equivalent* to the one used in the literature that we depicted on Figure D.1, if the IQ urn and the Random urn are being selected with equal probability. If the state is H, a ball indicating H is drawn with probability $0.5 \cdot 0.5 + 0.5 \cdot 1 = 0.75$, exactly the same as in Figure D.1. Similarly, Figure D.5 is equivalent to the information structure in Figure D.2 with the signal precision of 55%.

FIGURE D.4. Design developed in this paper (2 states).

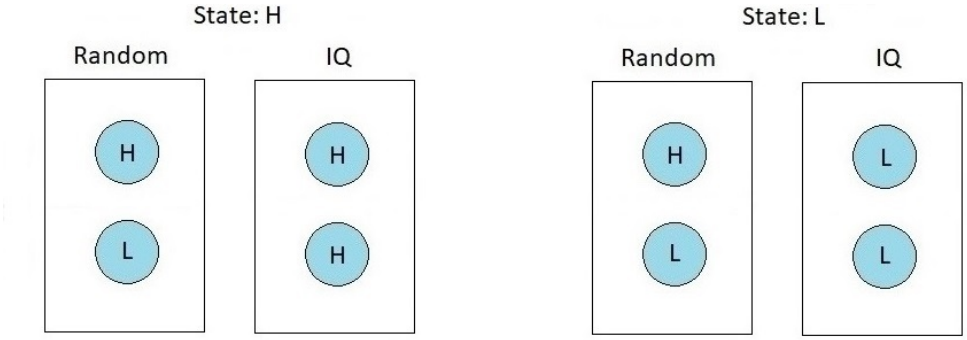
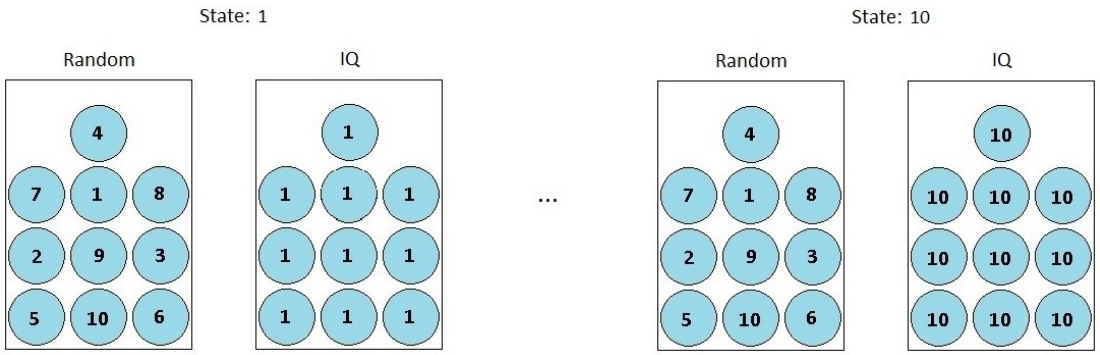


FIGURE D.5. Design developed in this paper (10 states).



[2] One could also distinguish the two urns along the dashed line in Figure D.2, with the Random urn containing all numbers except the one that indicates the state. This design, however, lacks the intuitive interpretation of "a random urn" from which *any number* can be drawn with *the same* probability, hence it might be more difficult to explain to the participants.

TABLE D.1. Literarure review: design comparison.

Other Work	**This Paper**	**Purpose**
1. Number of signals:		
– more than 1 signal Eil and Rao, 2011; Buser et al., 2018; Coutts, 2019; Zimmermann, 2020; Drobner and Goerg, 2021; Möbius et al., 2022.	– 1 signal	– separating reaction to signals from information aggregation.
– 1 signal Ertac, 2011; Schwardmann and Van der Weele, 2019; Drobner, 2022.		
2. State space, signal space, signal precision:		
– 2 states (above or below 50%; above or below 85% in Coutts, 2019), – 2 signal values, – signal precision: 67% Coutts, 2019; Drobner and Goerg, 2021; Drobner, 2022.	– 10 states (deciles of distribution) – 10 signal values – a signal is either perfectly informative or entirely uninformative (with equal probability.).	– richer state space and signal space to induce a stronger emotional reaction to a signal (based on the observation that it is more painful for subjects to be in the bottom 10% than in the bottom 50%). – by introducing signals that are perfectly informative or entirely uninformative (with equal probability), we reduce the compression effect described by Ambuehl and Li (2018).
– 2 states (above or below 50%) – 2 signal values – signal precision: 70% Buser et al., 2018.		
– 2 states (above or below 50%) – 2 signal values – signal precision: 75% Schwardmann and Van der Weele, 2019; Möbius et al., 2022.		
– 3 states (lower 20%, middle 60%, or upper 20%) – 2 signal values – perfectly informative but coarse signals Ertac, 2011.		
– 10 states (deciles of the distribution) – 2 signal values – signal precision depends on the state: 56%, 67%, 78%, 89% or 100%. Eil and Rao, 2011; Zimmermann, 2020.		

Other Work	This Paper	Purpose
3. Information structure and implementation:		
– information structure as in Figure D.1 – a signal is true or false with precision known to the subjects Buser et al., 2018; Coutts, 2019; Schwardmann and Van der Weele, 2019; Drobner and Goerg, 2021; Möbius et al., 2022. Drobner, 2022, uses the same information structure (Figure D.1), but the signal is a comparison with another subject.	– info structure as in Figure D.4. It is equivalent to the structure in Figure D.2 with a signal precision of 55%.	– it would not be possible to introduce richer state and signal space using any other information structure from the literature.
– information structure as in Figure D.3 – a signal is a pairwise comparison with another subject Eil and Rao, 2011; Zimmermann, 2020.		
– a signal is always true, but only reveals whether the subject is in the top or the bottom half of the distribution, and not precisely the state Ertac, 2011.		
4. Comparison group:		
– a group of 4 Schwardmann and Van der Weele, 2019; Drobner, 2022.	– 300 subjects	– a larger comparison group makes it more difficult to use reappraisal to lessen the impact of the negative signal (e.g., in the case of a group of four, one can easily attribute a negative signal to being assigned to a particularly strong pair of subjects). When there is another way of "explaining" a bad signal, there may be no need for (costly) belief distortion.
– a group of 8 Buser et al., 2018.		
– a group of 10 Eil and Rao, 2011; Ertac, 2011; Zimmermann, 2020.		
– a group larger than 10 Coutts, 2019; Drobner and Goerg, 2021; Möbius et al., 2022.		
5. Timing of revealing information:		
– In most of the papers mentioned above it is unclear whether and when subjects expected the resolution of uncertainty (see Drobner, 2022, for a comprehensive literature review). This problem was noticed and tested in a contemporaneous work of Drobner (2022).	– available online one week after the session	– to describe the behavior with a one-period model without dynamic concerns – to bring the design closer to the real-world situations: grades are rarely immediate, need to be checked etc.

APPENDIX E

Data Analysis: Payoffs

In this section, I look at the payoffs from the main task in the treatment and the control conditions. In both conditions, subjects were remunerated with "lottery tickets": a higher probability of receiving a large reward of 12 Euro. Decisions of participants in the treatment condition brought them, on average, 65.5% probability of receiving a large reward. At the same time, the average payoff taking all decisions in the control condition amounts to 78.2% probability of receiving a large reward (see Table E.1). However, the actual payoffs subjects received in the Control condition were much lower and not significantly different from the payoffs of participants in the Treatment condition. The discrepancy between the two is due to the fact that participants made much worse decisions when deciding about their actual rank than when deciding about a random number. This holds true both for the Treatment and the Control condition.

There are notable differences when comparing decisions in the Treatment and the Control condition separately for signals equal to one's rank and other signals, see Table E.2. When guessing about their actual rank, participants in the Treatment condition did better than subjects in the Control, although the difference of 8.4 percentage points is not statistically significant (p-value = 0.118). At the same time, subjects in the Control condition performed better when evaluating signals different from their true rank – the difference of 6.15 percentage points is significant at the 5% level (p-value = 0.024). The average payoffs in the two conditions mask considerable heterogeneity, which should be taken into account when making welfare comparisons.

TABLE E.1. Differences in payoffs from the main task.

				p-value	
	Treatment	**Control**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Payoffs (all decisions)	65.57%	78.16%	0.000	0.000	1.000
	(2.79)	(1.33)			
Payoffs (actual draw)	65.57%	64.96%	0.541	0.919	0.459
	(2.79)	(5.60)			
N	160	49			

TABLE E.2. Differences in payoffs when guessing one's rank.

				p-value	
	Treatment	**Control**	H_0: Diff < 0	Diff ≠ 0	Diff > 0
Payoffs (Signal = Rank)	53.51%	45.11%	0.882	0.237	0.118
	(4.41)	(5.57)			
N	73	49			
Payoffs (Signal ≠ Rank)	75.69%	81.84%	0.024	0.048	0.976
	(3.19)	(1.23)			
N	87	441			

E.1. Payoffs from Belief Elicitation I and II

In this section, I describe subjects' payoffs from the first and the second belief elicitation as well as the payoffs subjects would have gotten if they had rationally updated their beliefs about rank. First, let me compare the payoffs in the two conditions. The results are gathered in Table E.3. While signals moved subjects' beliefs in the Treatment condition ensuring a larger payoff, there is no significant difference in payoffs in the Control condition (which should not come as a surprise, since participants in the Control condition did not receive any new information).

TABLE E.3. Payoffs from Belief Elicitation I and II in the two conditions.

	Belief Elicitation I	**Belief Elicitation II**	**p-value** H_0: Diff < 0	Diff ≠ 0	Diff > 0
Treatment	47.35% (1.32)	51.02% (1.72)	0.046	0.092	0.954
Control	45.97% (2.39)	48.41% (2.41)	0.236	0.471	0.764

TABLE E.4. Payoffs from Belief Elicitation II and the rational update.

Payoff Elicitation II	**Payoff if rational**	Diff	**p-value** H_0: Diff < 0	Diff ≠ 0	Diff > 0
51.02% (1.72)	53.87% (2.37)	2.86%	0.97	0.06	0.03

Sample restricted to the subjects who assigned non-zero prior to the signal:

Payoff Elicitation II	**Payoff if rational**	Diff	**p-value** H_0: Diff < 0	Diff ≠ 0	Diff > 0
55.83% (2.37)	64.03% (2.31)	8.20%	1.00	0.00	0.00

Since only participants in Treatment received information that shifted their beliefs, let me focus only on these subjects. The difference in payoffs between the first and the second belief elicitation shows that even though the aggregate belief distribution seems to change little after the task (see Figure 2.10), the individual distributions changed in a way that guaranteed higher payoffs. Still, the payoffs would have been 2.86 percentage points higher (5.61% increase in relative terms), if participants had updated their beliefs rationally based on their prior belief distribution and the signal they received. The averages and the corresponding tests are gathered in Table E.4. However, the first difference was calculated including participants who assigned a prior probability of zero to the signal displayed on their screens. For those subjects, the rational posterior is assumed to be zero, and it reduces the average

difference. Indeed, for participants who assigned a non-zero prior probability to the signal displayed on-screen, the difference between the second belief elicitation and the rational benchmark is equal to 8.20 percentage points (14.69% increase in relative terms). It means that, for those participants, the payoff would have been almost 15% higher, if they had updated according to the Bayes rule.

Those subjects would also be better-off if they formed a rational posterior using their beliefs about the box. In Table E.5, I compare the payoffs from the second belief elicitation with the payoffs that participants would have gotten if they updated their beliefs about the rank in a way consistent with their decisions about the signal. The difference between payoffs in Belief Elicitation II and the Consistent Posterior is equal to 2.42 percentage points (4.74% increase in relative terms) and is significant at the 10% level. If we restrict the sample to the participants who assigned non-zero prior belief to the signal they received, the difference increases to 5.50 percentage points (9.85% increase in relative terms) and is significant at the 1% level.

TABLE E.5. Payoffs from Belief Elicitation II and consistent beliefs.

Payoff Elicitation II	**Payoff if consistent**	Diff	**p-value** H_0: Diff < 0	Diff ≠ 0	Diff > 0
51.02% (1.72)	53.44% (2.31)	2.42%	0.948	0.103	0.051

Sample restricted to the subjects who assigned non-zero prior to the signal:

Payoff Elicitation II	**Payoff if consistent**	Diff	**p-value** H_0: Diff < 0	Diff ≠ 0	Diff > 0
55.83% (2.37)	61.33% (3.70)	5.50%	0.993	0.013	0.007

Note: "Payoff if consistent" refers to the payoff from Belief Elicitation II if subjects formed beliefs consistent with their inference about the signal. "Restricted sample" includes only participants who received a signal to which they assigned a non-zero prior probability (87 subjects in the Treatment condition).

APPENDIX F

Information Acquisition

In this section, I describe the data from the very last part of the study. As I already mentioned in the main text, we informed participants that they will not learn the test result on the day of the experiment. They could obtain this information only one week later by clicking on a website that was created for the experiment. Every participant was given a sealed envelope with a personal link inside.[1] Under this link, one week after their session, they could find their rank in the IQ test, as well as the details of their payment. This personal information was not accessible to other participants, as only the person who knew the link (part of which was the participant's number and a four-digit code) could access it. The website was programmed in oTree and enabled us to collect information about participants who decided to check it.

Overall, 51% of all participants checked their links *even though* this part of the study was not incentivized (subjects did not get any money for it). There is no significant difference in information acquisition between the treatment and the control group (p-value = 0.962). While we cannot say for sure what motivated subjects to click or not (the reasons may range from simply losing the envelope to various motives described in the information avoidance literature, see Golman et al., 2017, for a literature review), we can check for individual traits that correlate with subjects' choices.

The results of simple regression analysis are gathered in Table F.1. The independent variable is an indicator variable taking value 1 if a subject decided to check the website. We observe that the lower the relative performance of a subject (the higher the rank) the lower the likelihood of checking the link. A person whose rank was Rank 1 will acquire information with 74% probability, while a subject ranked 10 – only with 29% chance. One possible explanation is that less cognitively able participants may be more likely to forget or lose the envelope, however, the next column in the regression shows that beliefs about the rank play a role. Participants with higher beliefs (lower perceived performance) tend to check the link

[1] Each envelope was placed in front of the subject, and its purpose was explained in the instructions. At the end of the experiment, research assistants reminded subjects not to forget the envelopes. The text inside informed subjects about the date and the type of information they can find under the link.

TABLE F.1. Performance, beliefs, and information acquisition.

	(1)	(2)	(3)	(4)	(5)
Median Belief			-0.05**		
			(0.02)		
Bias				0.05**	
				(0.02)	
Overconfident					0.15
					(0.09)
Rank		-0.05***	-0.04***	-0.09***	-0.07***
		(0.01)	(0.01)	(0.02)	(0.017)
Constant	0.51***	0.79***	0.95***	0.95***	0.80***
	(0.03)	(0.08)	(0.10)	(0.10)	(0.08)
Observations	209	209	209	209	209

Standard errors in parentheses

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

Note: The dependent variable is a dummy variable indicating whether or not a participant checked his IQ test result. The independent variable "Median Belief" refers to the rank that lies in the middle of the subject's prior belief distribution. "Bias" denotes the difference between the subject's actual rank and his median belief. It takes positive values for agents who overestimate their performance and negative values for those who underestimate it. The indicator variable "Overconfident" takes value 1 if the subject's bias is larger than zero.

less. An increase in the median belief by one rank translates to a 5 percentage point decrease in the probability of acquiring information.

To further investigate the link between the subject's rank, beliefs, and information acquisition, we look at the effect of the subject's bias, which we define as a difference between one's true rank and median belief (positive values indicate an overestimation of one's relative performance). The coefficient at the "Bias" variable is positive and significant, revealing that the larger the bias the more likely a subject is to acquire information if his bias has a positive sign (he tends to overestimate his performance) and the less likely if it has a negative sign (he underestimates his performance). In other words, overconfident subjects tend to seek information, while underconfident participants shy away from it. We obtain a qualitatively similar result if we regress our dependent variable on an indicator variable "Overconfident" taking value 1 if the subject's bias, as defined above, is larger than zero. Being overconfident is associated with a 15 percentage point higher probability of checking the link, controlling

TABLE F.2. Received signals, beliefs, and information acquisition.

	(1)	(2)	(3)	(4)	(5)	(6)
Signal Value	-0.026*	-0.015	-0.014			
	(0.014)	(0.014)	(0.014)			
Good Signal				0.002	-0.030	0.032
				(0.080)	(0.078)	(0.081)
Median Belief			-0.054**			-0.058**
			(0.022)			(0.024)
Rank		-0.048***	-0.038**		-0.052***	-0.040***
		(0.015)	(0.015)		(0.014)	(0.015)
Constant	0.647	0.860***	1.035***	0.505***	0.813***	0.974***
	(0.087)	(0.106)	(0.128)	(0.053)	(0.099)	(0.117)
Observations	160	160	160	160	160	160

Standard errors in parentheses

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

Note: The dependent variable is a dummy variable indicating whether or not a participant checked his IQ test result. The independent variable "Signal Value" denotes the value of the signal received (the number displayed onscreen). Higher values indicate worse signals. The independent variable "Median Belief" refers to the rank that lies in the middle of the subject's prior belief distribution. "Good Signal" is a dummy variable taking value 1 if a signal was better or equal to the subject's median belief.

for individual rank. However, the coefficient misses the conventional threshold for statistical significance (p-value = 0.117).

The relationship between the signal received and information acquisition is less clear, as presented in Table F.2. While the value of the signal seems to be related to our variable of interest as expected (the higher the rank displayed on-screen the lower the probability of checking the link), the effect is only significant at the 10% level (p-value = 0.071) and it loses significance once we control for subject's rank. There is also no significant effect of the signal valence, nor a positive or a negative surprise (defined as a difference between the signal and the median belief, not shown in the table). However, in all these cases, we can only analyze the behavior of participants in the treatment condition – those who received signals – reducing our sample size to 160. Any more complex relation between the subject's rank, beliefs, received signal, and information acquisition is unlikely to be found in the collected dataset.

TABLE F.3. Personality traits and information acquisition.

	(1)	(2)	(3)
Extraversion	0.003	0.008	0.009
	(0.01)	(0.01)	(0.01)
Conscientiousness	-0.015	-0.015	-0.012
	(0.01)	(0.01)	(0.01)
Openness	-0.007	-0.010	-0.011
	(0.01)	(0.01)	(0.01)
Neuroticism	-0.010	-0.008	-0.013
	(0.01)	(0.01)	(0.01)
Agreeableness	0.019	0.012	0.014
	(0.01)	(0.01)	(0.01)
Anxiety Trait			-0.005
			(0.01)
Anxiety State			0.002
			(0.01)
Rank		-0.049***	-0.048***
		(0.01)	(0.01)
Constant	0.589*	0.929***	1.089**
	(0.32)	(0.32)	(0.46)
Observations	209	209	209

Standard errors in parentheses

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

Note: The dependent variable is a dummy variable indicating whether or not a participant checked his IQ test result.

Last but not least, we look at correlations between information acquisition and personality traits, emotions experience during the task, and habitual use of emotion-regulation strategies. We report no significant correlation between any of the Big-5 personality traits nor STAI and information acquisition. In a regression including all personality and anxiety measures as independent variables, only Agreeableness comes close to being statistically significant (p-value = 0.107), with more agreeable individuals being more likely to check the link. However, its effect disappears if we control for the individual rank. In the second specification, we regress our variable of interest on the measures of achievement emotions and emotion regulation strategies, controlling for the subject's rank. Out of the eight achievement emotions, two are significantly correlated with information acquisition: anger and anxiety. Reporting

a 1 point higher feeling of anger on a 7-point Likert scale is associated with a 4.9 percentage point lower probability of checking the link (p-value = 0.063). At the same time, reporting a 1 point higher feeling of anxiety is related to a 9.6 percentage point higher probability of acquiring information (p-value = 0.040). Neither reappraisal nor suppression is correlated with information acquisition.

TABLE F.4. Achievement emotions and information acquisition.

	(1)	(2)	(3)
Enjoyment	0.004	-0.005	-0.006
	(0.02)	(0.02)	(0.02)
Hope	0.006	-0.001	-0.003
	(0.03)	(0.03)	(0.03)
Pride	0.044	0.050	0.050
	(0.03)	(0.03)	(0.03)
Relief	0.046	0.044	0.043
	(0.03)	(0.03)	(0.03)
Anger	-0.043	-0.049*	-0.047*
	(0.03)	(0.03)	(0.03)
Anxiety	0.089*	0.096**	0.095**
	(0.05)	(0.05)	(0.05)
Shame	-0.024	-0.004	-0.004
	(0.03)	(0.03)	(0.03)
Hopelessness	0.019	0.006	0.007
	(0.04)	(0.04)	(0.04)
Reappraisal			0.022
			(0.03)
Supression			-0.020
			(0.04)
Rank		-0.048***	-0.048***
		(0.01)	(0.01)
Observations	209	209	209

Standard errors in parentheses

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

Note: The dependent variable is a dummy variable indicating whether or not a participant checked his IQ test result. The independent variables denote survey measures of the achievement emotions. "Reapp" and "Supres" denote the two emotion regulation strategies: reappraisal and supression. All specifications include a constant (omitted in the table).

CHAPTER 3

Hope for the Best, Prepare for the Worst: Signal Anticipation and Ex-ante Belief Manipulation

This paper investigates how people form beliefs about parameters relevant to their self-esteem. In particular, I look at the formation of beliefs about one's cognitive ability (intelligence). Beliefs about cognitive ability drive many economic decisions, e.g., the choice of career path, the decision to become an entrepreneur, the formation of expectations about future earnings, and consumption-savings decisions. A large body of work has shown that people tend to overestimate their abilities – a bias known as overconfidence. While many papers document the bias and its consequences, there are still open questions about the way it arises.[1] The mechanism that I examine in this paper concerns systematic distortions in the process of belief formation.

Most economists conceptualize belief formation as *belief updating* assuming that an agent starts with a well-defined prior and incorporates new information using the Bayes' rule. Although it is a good approximation to reality in some contexts, the Bayesian model seems to be less adequate in others. One such example is a situation in which the decision-maker has a strong preference over the states of the world, as in the case of ego-relevant characteristics. Several studies demonstrated that people significantly deviate from the Bayes' rule when forming beliefs about their intelligence or beauty (Benjamin, 2019). Yet, the direction of the effect and its magnitude differ across studies.[2]

An alternative modeling approach, developed to explain unrealistically optimistic beliefs in the financial domain (Brunnermeier and Parker, 2005), acknowledges additional motives to form specific beliefs and allows an agent to directly choose her beliefs. The agent maximizes her consumption utility and utility derived from beliefs. When selecting her beliefs,

[1] For a comprehensive review of the literature on the origins of overconfidence I refer the reader to Burks et al. (2013), Bénabou and Tirole (2016), and Huffman et al. (2022).

[2] Some authors found positive asymmetry in updating (Eil and Rao, 2011; Kozakiewicz, 2020; Drobner and Goerg, 2021; Möbius et al., 2022), others found no asymmetry (Buser et al., 2018; Schwardmann and Van der Weele, 2019; Zimmermann, 2020), or even negative asymmetry (Ertac, 2011; Coutts, 2019).

she faces a trade-off between belief accuracy – necessary to take a profit-maximizing action – and their desirability. While some models predict overly optimistic beliefs (Brunnermeier and Parker, 2005; Bracha and Brown, 2012; Caplin and J. V. Leahy, 2019), others describe conditions under which the agent adopts pessimistic beliefs (Gollier and Muermann, 2010; Macera, 2014). Our theoretical framework builds upon the latter. I assume that the agent's utility is reference-dependent (Kőszegi and Rabin, 2006), hence I introduce incentives to adopt overly pessimistic beliefs. The idea of reference-dependent utility from beliefs was developed by Kőszegi and Rabin (2009). However, their approach is conceptually different, as beliefs are not subject to a choice but are rationally formed based on credible plans of future behavior.[3]

The intuition behind the model presented in this paper is that people can "prepare themselves" for the arrival of new information by adopting overly pessimistic beliefs. The prior is not fixed but can be manipulated depending on subjects' expectations over the upcoming signal. While the phenomenon of "expectation management" (also referred to as "bracing") is well-established in the psychological literature (see, for example, Shepperd et al., 1996; Carroll et al., 2006; Sweeny et al., 2006; Sweeny and Krizan, 2013), none of the studies look at a shift in beliefs before a partial information revelation.[4]

Why would people brace themselves before receiving a noisy signal? If an agent expects a signal to move his beliefs and bring him (dis)utility, he can lower his prior to 1) reduce the painful downward shift in beliefs after a negative signal and 2) increase the pleasant upward shift after a positive signal. The crucial assumption is that the utility from beliefs is reference-dependent, and the current belief level serves as a reference point. Then, the agent can increase his expected utility by shifting it – adopting a more pessimistic belief. However, there is a trade-off: lower beliefs imply lower utility from beliefs right now. The solution to the problem depends on the agent's attitudes towards gains and losses in belief-based utility.

The model makes several predictions. First of all, when an agent expects to receive a noisy signal, he chooses beliefs that are lower (more pessimistic) than the beliefs he would choose

[3] In the rational-expectations models, an agent derives utility from beliefs that are consistent with his actions in equilibrium (Caplin and J. Leahy, 2001; Kőszegi and Rabin, 2009; Kőszegi, 2010).

[4] I postpone explaining why it is important to look at a noisy signal (partial information revelation) instead of the outcome (full information revelation) until I describe the experimental design. The robust finding in the case of a full information revelation (the state is fully revealed at the end of the waiting period) is that beliefs tend to follow a downward path, reaching the lowest point right before learning the state. One would expect a similar shift in beliefs, albeit with a lower magnitude, before a signal.

if he did not expect any new information. In the latter, the incentives to hedge are mitigated, making the motive to maintain a positive self-view the dominant one. Second, more loss-averse subjects will lower their beliefs to a larger extent. Intuitively, an agent who experiences losses in belief-based utility as more painful will react more strongly to the prospect of a signal. Third, the negative effect of loss aversion will be mitigated for agents with higher ability. The probability of receiving a "bad" signal is lower for high-ability agents, resulting in a lower weight placed on the loss component.

The model also predicts that agents who are non-loss-averse and have low ability will overestimate their ability, and those characterized by aversion to losses and high ability will underestimate it. The bias of agents who overestimate their ability (overconfident agents) is expected to be larger than the bias of underconfident individuals. Since beliefs are driven by loss aversion, so is the agent's bias: it will be decreasing (increasing) in the loss aversion parameter for overconfident (underconfident) agents.[5]

In order to test the model predictions, I designed a simple experiment. First, subjects solved an IQ test that allowed us to measure their cognitive ability.[6] After the test, they reported subjective probability that their test score placed them in the i^{th} decile of the distribution of IQ test scores obtained from a large group of former participants. We informed subjects that their results will be available to them online, one week after their session.[7] I delayed the full information revelation to minimize confounding factors (see footnote 10) and focus on the trade-off captured by the model.

I introduced two experimental conditions that varied with respect to the timing of information given to subjects. In the Treatment condition, we informed participants *before* belief elicitation that, later in the session, they will receive a noisy signal about their relative performance. They were familiarized with the signal structure and instructed on how the signal will be drawn. In the Control condition, participants received the same information but only after

[5]Here, I refer to the absolute bias (defined as the absolute difference between one's beliefs and ability).

[6]I use cognitive ability as the ego-relevant parameter of choice for several reasons. First, intelligence is a personal characteristic that people deeply care about. It is particularly relevant in a university setting (for this reason, I run in-person sessions at the University of Bonn). Importantly, there are established methods to assess it, providing us with a measure that is reliable, valid, and easy to obtain.

[7]Each participant was given a link to an anonymous website on which he could see, one week after the session, his (and only his) IQ test result, position in the distribution, and payment details. Moreover, we prevented subjects from inferring their scores from the final payoffs by adding up their earnings from different tasks. Subjects were informed that they will be paid at the end of the session, but the details of their payments (how much they earned in each task) will be available to them only one week later.

they reported their prior beliefs. At the moment of belief elicitation, they had no information about a signal. In both conditions, subjects completed additional tasks designed to obtain two independent measures of loss aversion.[8] The first measure was based on a hypothetical scenario. Before the IQ test, participants were told to imagine that they took an important exam and are about to receive information about its outcome – whether their result was better or worse than they expected. The instructions made it clear that the signal will only shift their beliefs from 50% to 70% and it will not fully reveal the state. Then, subjects answered two hypothetical questions intended to assess their utility before receiving a signal. Those who indicated higher responses on a 9-point Likert scale were classified as more loss-averse. The second measure of loss aversion was based on the willingness to pay for a signal conditional on its realization (unlike the first measure, the second measure was incentivized). Participants were told that, before displaying a signal, the computer program will check the draw and, depending on their choice for this realization, display it to them or not. They filled in 10 price lists, one for each signal, knowing that one of them (the one corresponding to the actual draw) would be implemented.

The experiment was conducted in the summer of 2022 at BonnEconLab – the decision lab under the patronage of the University of Bonn. I collected data from 234 participants, mostly university students. The sessions lasted around 85 minutes and the average earnings were equal to 19 euros. I test the model predictions using the mean of individual belief distribution reported in the main task (the results are the same when I use the median or any quartile instead). On average, subjects revealed the mean belief above the 6^{th} decile of the distribution – the belief that was significantly higher than the average actual position. The average treatment effect is negative, as predicted by the model, and equal to -0.25 (one-fourth of a decile). Unfortunately, it is not significant at any acceptable level (p-value of one-sided t-test $= 0.287$).

In order to test the second prediction of the model, I examine the relationship between participants' beliefs and their loss attitudes. The first measure of loss aversion is characterized by a distribution that is close to symmetric, with a mean of 4.84 and a standard deviation of 1.76. The median is equal to 5, which is also the middle value on the Likert scale used in

[8] Goette et al. (2019) prove the importance of controlling for gain-loss attitudes when testing models of reference-dependent preferences. They also show that loss attitudes are not correlated across domains, hence one could not use the aversion to losses in the financial domain in place of belief-based utility.

this task. The results of a regression analysis reveal heterogeneous treatment effects: more loss-averse participants adopted more pessimistic beliefs in the Treatment condition. The coefficient at the interaction between the Treatment dummy and the loss aversion parameter is equal to -0.271 and is significant at the 5% level (p-value of one-sided t-test = 0.036). There is no correlation between beliefs and the loss aversion parameter in the Control condition. Both results are in line with the model predictions. Moreover, the effect of loss aversion on beliefs is mitigated for participants with higher ability (higher position in the IQ test score distribution). As predicted, the coefficient at the interaction of the loss aversion parameter and ability is positive and highly significant (p-value of one-sided t-test = 0.022). Lastly, I estimate a saturated model, in which the mean belief is regressed on the treatment dummy, the loss aversion parameter, subject's ability, and their interactions. While the theory predicts that the coefficient at triple interaction should be positive, I cannot confirm this in the data: the estimated coefficient is equal to 0.001 and not significant.

The attempt to obtain a second measure of loss aversion was less successful. Most participants were not willing to forgo as little as 10 cents to lower the probability of receiving a signal. Around 80% of all decisions were payoff-maximizing and differed only in participant's decision to see or not to see the signal at the point of monetary indifference. This prevented me from retrieving loss aversion parameters from the choice data. Instead, I coarsely classify subjects as "loss-averse" and "non-loss-averse" based on their decisions regarding whether or not to see the worst signals. I test the theory using an indicator variable "Loss Aversion". The effects are 20-50% lower and much noisier, however, it is reassuring to see that all estimates go in the predicted direction.

Last but not least, I examine the data on subjects' bias. I define a person as overconfident (underconfident) if their mean belief was higher (lower) than their position in the test score distribution. 60% of participants were classified as overconfident, and 40% as underconfident. As predicted by the model, the majority of the low-ability, non-lost-averse subjects were overconfident (the fraction amounts to 94%). At the same time, the fraction of underconfident among high-ability and loss-averse subjects was 74%. However, the relationship between confidence and ability can arise mechanically in any setup with relative performance: low-ability subjects are less likely to be underconfident because their beliefs are bounded from below. To address this confound, I test whether the respective probabilities are higher *than they would be* if subjects' beliefs were assigned randomly. I simulate the data

by randomly drawing ability and loss aversion parameters from their empirical distributions, and beliefs from the uniform distribution. A comparison with coefficients estimated on simulated data strongly confirms the hypothesis for overconfident, but not for underconfident agents.

Moreover, the absolute bias of overconfident subjects is 80% higher than the bias of underconfident participants. The difference is significant at the 1% level, providing strong evidence for the model. It is also much larger than any difference that emerged in the simulation: the estimated coefficient is equal to 1.155, whereas the value at the 99^{th} percentile of the distribution of simulated coefficients is lower than 0.5. For overconfident subjects, bias is driven by the loss aversion parameter as predicted: a higher aversion to losses results in a lower bias. I do not find the effect for underconfident subjects, which is not surprising – underconfident participants tend to be of higher ability, and the model predicts that, for high-ability agents, the effect of loss aversion is mitigated.

All things considered, the collected data is in line with the theory. Although the treatment manipulation is rather subtle, it provides evidence for the mechanism of the model, that is, the way loss aversion drives subjects' beliefs in the two conditions. At the same time, all other estimates have the predicted signs (including those that cannot be accepted with sufficient confidence). Moreover, the evidence on over- and underconfidence provides further support for the theory. These results are encouraging; they suggest that, with additional data, one can confirm all predictions of the model.

This paper contributes to the empirical literature on belief formation. In particular, the formation of beliefs when the decision-maker derives utility from his convictions.[9] To the best of my knowledge, it is the first paper to directly test a model of belief choice with reference-dependent utility. Consequently, there is no study measuring gain-loss attitudes towards signals. A few papers looked at beliefs formed before learning the final outcome.[10] This includes an experimental study by Van Dijk et al. (2003), which provides evidence on lowering

[9]Behavioral economics has long recognized anticipatory feelings – emotions such as anxiety or hope, arising from *beliefs* about the future – as drivers of human behavior. One of the first papers incorporating anticipatory emotions into an economic model were Akerlof and Dickens (1982) and Bell (1985).

[10]A set-up in which agents observe the outcome instead of receiving a noisy signal has several disadvantages. First, the willingness to pay to get to know the result might reflect factors different from gain-loss attitudes, e.g., curiosity or a desire to end the painful waiting period. These are less of a concern when the state is not fully revealed. Second, I aim to examine the functional form of belief-based utility. Providing evidence on the special case of beliefs shifting to certainty would be less informative about the general formulation. Lastly, subjects might respond to learning their IQ by manipulating their beliefs about test accuracy (Kozakiewicz, 2020). In this case, the decision to lower one's beliefs ex-ante would depend on the expected manipulation in the second period. By

one's beliefs before learning a test score.[11] A more recent work by Drobner (2022) shows that people update more optimistically when they know that the outcome will not be revealed to them.[12] Belief choice in a different setting was also studied in Engelmann et al. (2019). The authors provide evidence on "wishful thinking" (Caplin and J. V. Leahy, 2019), a tendency to adopt optimistic beliefs when expecting an unpleasant outcome. Contrary to my work, they study the formation of beliefs about a physically painful experience (an electric shock) and do not consider reference dependence. Other experimental literature focused so far on documenting deviations from the rational update (Eil and Rao, 2011; Ertac, 2011; Coutts, 2019; Kozakiewicz, 2020; Drobner and Goerg, 2021; Möbius et al., 2022), with more recent papers unraveling the factors driving these deviations. My results add to this line of research as they suggest that the inconsistent findings in the updating literature might be partly due to differences in subjects' expectations and gain-loss attitudes between the samples.

More broadly, my work contributes to the literature on motivated reasoning (see Bénabou and Tirole, 2016, for a review of the literature). This strand of research describes various strategies that people use to bias their beliefs to achieve certain goals. I add to this literature by providing evidence on the functional form of belief-based utility, which can be used further to describe processes behind phenomena such as asymmetric updating or information avoidance. Moreover, this is the first paper to establish a direct link between gain-loss attitudes and overconfidence. I propose and test a new mechanism that gives rise to overconfidence, complementing the literature that investigates its origins (see, for example, Burks et al., 2013; Schwardmann and Van der Weele, 2019; Huffman et al., 2022). The paper proceeds as follows. In the next section, I describe the model. The experimental design is outlined in Section 3.2 In Section 3.3, I formulate the hypotheses and explain in detail how I test them in the data. The data analysis and results can be found in Section 3.4. The last section concludes.

3.1. Model

An agent is learning about an unknown, ego-relevant state of the world $\omega \in \{H, L\}$. Let us interpret the state to be the level of cognitive ability, either high or low. The agent has a prior

postponing the full information revelation, we make this strategy less salient, so one can focus on the trade-off described in the model.

[11] A result very much in line with the psychological literature on bracing (see footnote 4).

[12] In line with the updating literature, Drobner (2022) focuses on changes in beliefs and not the belief choice per se. Although belief updating is no less important, I believe that the first step – how an agent forms her prior – is necessary to fully understand the dynamics of belief formation.

belief about his ability being high: p_0. There are two periods $t \in \{0,1\}$. At the beginning of the last period, Period 1, he receives a signal about the state $s \in \{H,L\}$ with known precision, and forms a posterior belief p_1. The agent updates his beliefs according to Bayes' rule and does not suffer from any information-processing bias. However, the agent derives utility from his beliefs about his cognitive ability. At any point in time, his utility from beliefs is $u(p_t) = p_t u_H + (1-p_t)u_L$.[13] Because of belief-based utility, the agent has incentives to manipulate his beliefs, forming a new prior $\tilde{p}_0$, which then enters his utility function.

The manipulation of beliefs in Period 0 comes at a cost. First of all, there is a cost of belief distortion, $\frac{\gamma}{2}(\tilde{p}_0 - p_0)^2$, that is a function of the distance to the true belief p_0.[14] Moreover, the agent knows that he will receive a signal in Period 1 and, although he values his beliefs in Period 0, he dislikes being negatively surprised by the signal. In Period 1, he experiences 1) utility from updated beliefs, and 2) gain-loss utility that is captured by a function $\mu(\cdot)$ and stems from a comparison between the utility induced by the posterior belief p_1 and the utility from prior beliefs.

In Period 0, the agent chooses $\tilde{p}_0$ to maximize the following:

$$\begin{aligned} U_0 = u(\tilde{p}_0) + P(s=H|p_0) \underbrace{\Big[u(p_1^H) + \mu\Big(u(p_1^H) - u(\tilde{p}_0)\Big)\Big]}_{\text{the utility in Period 1 after a signal } s=H} + \\ + P(s=L|p_0) \underbrace{\Big[u(p_1^L) + \mu\Big(u(p_1^L) - u(\tilde{p}_0)\Big)\Big]}_{\text{the utility in Period 1 after a signal } s=L} - \frac{\gamma}{2}(\tilde{p}_0 - p_0)^2, \end{aligned} \tag{3.1}$$

where $u(\tilde{p}_0) = \tilde{p}_0 u_H + (1-\tilde{p}_0)u_L$ is utility from beliefs manipulated in Period 0, $u(p_1^H) = p_1^H u_H + (1-p_1^H)u_L$ and $u(p_1^L) = p_1^L u_H + (1-p_1^L)u_L$ denote the utility from unmanipulated beliefs shifted by a signal $s=H$ and $s=L$, respectively. $P(s=H|p_0)$ denotes the probability of receiving a "good" signal $s=H$ given the prior belief p_0. The probability of receiving a "bad" signal is $P(s=L|p_0) = 1 - P(s=H|p_0)$. Our model is very similar to Gollier and Muermann (2010), as the agent chooses subjective probabilities facing a trade-off between ex-ante

[13] The current belief-based utility could be also interpreted as anticipatory utility about future consumption. In this interpretation, u_H (u_L) is the utility from being a high (low) type experienced in the future, e.g., a high consumption level after getting a well-paid job as a high type.

[14] The parameter γ captures the costs that are unrelated to the gain-loss component of the utility function. They can include, for example, the agent's cognitive limitations or how far he can move his beliefs without losing confidence in them. Although in principle these costs can differ depending on the direction of belief manipulation, I assume that they are symmetrical.

anticipatory utility and ex-post disappointment, but I add a quadratic cost of belief manipulation as in Engelmann et al. (2019). In order to evaluate the utility in Period 1, I use the gain-loss utility function as in Kőszegi and Rabin (2009): $\mu(x) = \eta x$ for the arguments in the domain of gains, and $\mu(x) = \lambda\eta x$ in the domain of losses, with $\eta > 0$ and $\lambda > 0$. The parameter η denotes the weight placed on the gain-loss component and λ is the loss aversion parameter. I set the reference point to the agent's beliefs from the previous period. In what follows, I focus on the case when an agent is in the gain domain after a good signal, and in the loss domain after a bad signal.

Importantly, I do not allow an agent to freely choose *every* belief-based aspect of the problem. In doing so, I follow the literature (Brunnermeier and Parker, 2005; Gollier and Muermann, 2010; Macera, 2014; Caplin and J. V. Leahy, 2019). I assume that 1) the agent is not distorting the probability of receiving a "good" signal $P(s = H|p_0)$, and 2) the posterior probabilities p_1^H and p_1^L follow from the true prior p_0.[15] The second point embodies the observation that, while people tend to perceive themselves in an unrealistically positive way, they do not seem to act upon these beliefs all the time. While the assumption that people can hold more than one belief might seem unusual, there is recent experimental evidence that people hold multiple prior beliefs (Abdellaoui et al., 2021). One can also view this assumption as a modeling technique, conceptually similar to describing time inconsistency with a dual-self model (Fudenberg and Levine, 2006). I use it to describe an internal, subconscious process of coming to the belief about one's ability.

The first-order condition gives us the following formula for the optimal prior:

$$\tilde{p}_0^{\,*} = p_0 + \frac{1}{\gamma}\Big(1 - P(s = H|p_0)\,\eta - P(s = L|p_0)\,\lambda\eta\Big)(u_H - u_L). \tag{3.2}$$

Expecting to receive a signal creates an additional incentive for the agent to manipulate his beliefs to 1) reduce the disutility from being negatively surprised, and 2) increase the utility from a positive surprise. These two effects are weighted with the objective probabilities of receiving a signal of each type. Additionally, the loss in belief-based utility might be more pronounced than a similar gain; this idea is captured by the loss aversion parameter λ. The two motives pull the beliefs downwards, counteracting the incentive to adopt overly optimistic beliefs to derive higher utility in Period 0.

[15] The next step would be to develop a dynamic model, in which the agent can manipulate his belief in Period 1 and takes it into account when choosing beliefs in Period 0.

3.1.1. No information about a signal. How one can describe an agent who does not know that he will receive a signal? The answer to this question is relevant for our experimental design. I model this situation in the following way. When choosing a new belief $\hat{p}_0$, the agent knows that he will keep this belief in Period 1 and derive belief-based utility from it. I assume that receiving a signal is not a zero-probability event – the agent assigns a probability ϵ to it. He expects to keep the manipulated belief with probability $(1-\epsilon)$ and, with probability ϵ, to receive a signal that will shift his beliefs.[16] He chooses $\hat{p}_0$ to maximize:

$$U_0 = u(\tilde{p}_0) + \underbrace{(1-\epsilon)\ u(\tilde{p}_0) + \ \epsilon\ P(s=H|p_0)\left[u(p_1^H)\ +\ \mu\Big(u(p_1^H) - u(\tilde{p}_0)\Big)\right] +}_{\text{the utility in Period 1}}$$

(3.3)

$$\underbrace{+\ \epsilon\ P(s=L|p_0)\left[u(p_1^L)\ +\ \mu\Big(u(p_1^L) - u(\tilde{p}_0)\Big)\right]}_{\text{the utility in Period 1}} - \frac{\gamma}{2}(\tilde{p}_0 - p_0)^2.$$

The first-order condition gives us the following formula for the optimal prior:

$$\hat{p}_0^{\,*} \ =\ p_0\ +\ \frac{1}{\gamma}\Big(2-\epsilon-\epsilon P(s=H|p_0)\eta-\epsilon P(s=L|p_0)\lambda\eta\Big)(u_H-u_L). \tag{3.4}$$

For $\lambda > 0$, $\eta = 1$, and $\epsilon \in (0,1)$, we have: $\tilde{p}_0^{\,*} < \hat{p}_0^{\,*}$. The optimal prior chosen when expecting a signal is always lower than the prior chosen not knowing about a signal.

Prediction 1

Optimal prior chosen knowing about an upcoming signal is lower (more pessimistic) than the optimal prior chosen when not knowing about a signal.

In the case of no information, chances of being disappointed (or elated) by a signal are diminished. The incentives to lower one's beliefs in anticipation of future shifts are less pronounced, making the utility from beliefs *right now* a dominating factor. The resulting belief is more optimistic than the belief adopted when the agent expects to receive a signal.

[16] It implies that the agent knows the signal structure and how a signal will shift his beliefs bringing belief-based utility. While these assumptions might appear unrealistic at first sight, one can argue that the agent could have encountered similar situations in the past that left a lasting impression of what kind of signals he receives and how they make him feel. For example, a student might not expect an unannounced test but he does assign ϵ probability to it and, having solved similar tests in the past, has some idea of what kind of signals about his knowledge or ability it will generate.

3.1.2. Comparative statics: loss aversion. The optimal solution (2) depends on the loss-aversion parameter:

$$\frac{\partial \tilde{p}_0^*}{\partial \lambda} = -\frac{1}{\gamma}\,\eta\, P(s=L|p_0)(u_H - u_L) < 0. \tag{3.5}$$

The more loss-averse the agent is, the lower the prior he chooses in Period 0. Anticipating a painful loss, a loss-averse agent tries to mitigate its impact by setting a lower reference point. This leads us to the following testable prediction:

Prediction 2.1

A more loss-averse agent will adopt a lower prior belief $\tilde{p}_0^*$ compared to a less loss-averse agent with the same unmanipulated belief p_0.

Note that, even when not expecting a signal, the optimal prior depends on λ:

$$\frac{\partial \hat{p}_0^*}{\partial \lambda} = -\frac{1}{\gamma}\,\epsilon\,\eta\, P(s=L|p_0)(u_H - u_L) < 0. \tag{3.6}$$

Since $\epsilon \in (0,1)$, we have:

$$\frac{\partial \tilde{p}_0^*}{\partial \lambda} < \frac{\partial \hat{p}_0^*}{\partial \lambda}. \tag{3.7}$$

The slope is steeper in (5), meaning that the optimal prior is more responsive to changes in λ when the agent knows about a signal.

Prediction 2.2

The loss aversion parameter λ has a more negative effect on the optimal prior $\tilde{p}_0^*$ when expecting a signal compared to the effect in the case of no information.

When the agent expects a signal, the weight placed on the gain-loss component is higher than in the no-information scenario, enhancing the effect of loss aversion.

3.1.3. Comparative statics: unmanipulated beliefs. The probability of receiving a high signal, $P(s=H|p_0)$, depends on a signal structure as well as the unmanipulated belief p_0. Let us consider the following signal structure: if the state is H, the agent receives a signal H with probability c, $c > 0.5$, and a signal L with probability $(1-c)$. If the state is L, the agent receives a

signal L with the same probability c, and a signal H with probability $(1-c)$. The probability of receiving a high signal is $P(s=H|p_0)=(2c-1)p_0+1-c$. Rewriting (2) and taking the derivative with respect to p_0 gives us the following condition:

$$\frac{\partial \tilde{p}_0^*}{\partial p_0} = 1 - \frac{1}{\gamma}(2c-1)\eta(1-\lambda)(u_H - u_L). \tag{3.8}$$

For $\lambda > 1$, (8) is always larger than zero. The effect of unmanipulated prior on the optimal belief $\tilde{p}_0^*$ is positive. For $\lambda < 1$, the sign is positive for loss aversion parameters higher than $\bar{\lambda} = 1 - \gamma \,/\, [(2c-1)\eta(u_H - u_L)]$ and negative otherwise:[17]

$$\begin{aligned}
&\frac{\partial \tilde{p}_0^*}{\partial p_0} < 0 && \text{for} \quad \lambda < 1 \quad \text{and} \quad \lambda < \bar{\lambda}, \\
&\frac{\partial \tilde{p}_0^*}{\partial p_0} > 0 && \text{for} \quad \lambda < 1 \quad \text{and} \quad \lambda > \bar{\lambda}, \\
&\frac{\partial \tilde{p}_0^*}{\partial p_0} > 0 && \text{for} \quad \lambda > 1.
\end{aligned} \tag{3.9}$$

Prediction 3.1

For loss-averse agents, an increase in the unmanipulated belief p_0 has a positive effect on the optimal belief $\tilde{p}_0^*$. For non-loss-averse agents, it depends on parameters λ, γ, c, u_H, and u_L in a way described by (9).

The intuition behind Prediction 3.1 is the following. As p_0 increases, the agent can increase the manipulated belief staying at the same cost curve. He can change $\tilde{p}_0$ to the same extent, which is the meaning behind the "1" in (3.8). Manipulating one's beliefs upwards is desirable, as it increases the utility in Period 0. At the same time, an increase in $\tilde{p}_0$ has a negative effect on utility due to the gain-loss component. It lowers the gains after $s = H$ and increases the losses after $s = L$. However, an increase in p_0 also affects the posterior beliefs as well as the probabilities of receiving a "good" and a "bad" signal. It shifts $P(s=H|p_0)$ upwards, exacerbating the negative effect of an increase in $\tilde{p}_0$ from the gain component, and it shifts $P(s=L|p_0)$ downwards, mitigating the negative effect from losses. For loss-averse agents, the latter receives a higher weight as $\lambda > 1$. The net effect of increasing $\tilde{p}_0$ is positive, so the agent will adopt a higher belief. If the loss aversion parameter λ is below 1, the mitigating effect

[17] The threshold value $\bar{\lambda}$ can be found by equating (8) to zero.

receives a lower weight. Adopting a higher $\tilde{p}_0$ might be profitable, depending on the relation between λ and other parameters.

Secondly, I compare the scenario when the agent expects to receive a signal to the no-information case. In the later, the partial derivative with respect to p_0,

$$\frac{\partial \hat{p}_0^*}{\partial p_0} = 1 - \epsilon \frac{1}{\gamma}(2c-1)\eta(1-\lambda)(u_H - u_L), \tag{3.10}$$

is positive for non-loss-averse agents. For loss-averse agents, it follows conditions similar to (8), with a threshold $\bar{\lambda}^{\epsilon}$ derived from (9). Comparing the two scenarios, we obtain the following conditions:

$$\begin{aligned} &\frac{\partial \tilde{p}_0^*}{\partial p_0} < \frac{\partial \hat{p}_0^*}{\partial p_0} \quad \text{for} && \lambda < 1, \\ &\frac{\partial \tilde{p}_0^*}{\partial p_0} > \frac{\partial \hat{p}_0^*}{\partial p_0} \quad \text{for} && \lambda > 1. \end{aligned} \tag{3.11}$$

Prediction 3.2

For loss-averse agents, the effect of the unmanipulated belief p_0 on the optimal belief $\tilde{p}_0^*$ is larger when an agent expects to get a signal compared to the case when he does not expect a signal. For non-loss-averse agents, the opposite is true.

As the gain-loss component loses its importance in the no-information case, the effect of p_0 on $\tilde{p}_0$ is reduced. When an increase in p_0 causes an increase in $\tilde{p}_0^*$ (the case of loss-averse agents), the belief in the no-information scenario is lower than the belief chosen when expecting a signal. When an increase in p_0 causes a decrease in $\tilde{p}_0^*$ (non-loss-averse agents), the opposite is true. (11) also holds when derivatives have opposite signs.

3.1.4. Comparative statics: loss aversion and unmanipulated beliefs. Lastly, I examine interaction effects between the loss aversion parameter and unmanipulated beliefs. I look at the sign of:

$$\frac{\partial \partial \tilde{p}_0^*}{\partial \lambda \partial p_0} = \frac{1}{\gamma}\eta(2c-1)(u_H - u_L), \tag{3.12}$$

which is a mixed derivative with respect to λ and p_0. Since $c > \frac{1}{2}$ and $u_H > u_L$, the derivative has a positive sign. An increase in the unmanipulated belief p_0 has a positive effect on the relation between loss aversion and chosen beliefs. In other words, increasing p_0 counteracts the negative effect of loss aversion. As a result, the loss aversion parameter will have a lesser effect on agents with higher p_0. I conclude that:

Prediction 4.1

An increase in the unmanipulated belief p_0 lessens the negative effect of loss aversion on the optimal belief $\tilde{p}_0^*$.

Moreover, the effect is more pronounced when the agent expects to receive a signal. By comparing (12) to the analogous mixed partial derivative based on (10), we get:

$$\frac{\partial\partial \tilde{p}_0^*}{\partial\lambda\partial p_0} > \frac{\partial\partial \hat{p}_0^*}{\partial\lambda\partial p_0}. \tag{3.13}$$

One can state the following prediction:

Prediction 4.2

An increase in the unmanipulated belief p_0 lessens the negative effect of loss aversion on the optimal belief $\tilde{p}_0^*$ to a larger extent when the agent expects to receive a signal compared to the case when he does not expect a signal.

3.1.5. Boundary cases. It is important to consider two limiting cases: when the unmanipulated belief is equal to zero (being a low type with probability one) and when it is equal to one (being a high type with probability one). In the first case, the optimal belief is described by:

$$\tilde{p}_0^* = \frac{1}{\gamma}\, c\,(1-\lambda)(u_H - u_L). \tag{3.14}$$

The optimal belief cannot be negative, so an agent with $\lambda > 1$ chooses the boundary solution $\tilde{p}_0^* = 0$ (regardless of whether he expects a signal or not). The prediction $\tilde{p}_0^* < \hat{p}_0^*$ cannot be made for those agents. For non-loss-averse agents ($\lambda < 1$), the inequality $\tilde{p}_0^* < \hat{p}_0^*$ holds – they adopt more optimistic beliefs when expecting a signal.

In the second case, $p_0 = 1$, the optimal belief is equal to:

$$\tilde{p}_0^{\,*} = 1 + \frac{1}{\gamma}(1-c)(1-\lambda)(u_H - u_L). \tag{3.15}$$

Since $0.5 < c < 1$ and $\tilde{p}_0^{\,*}$ cannot be larger than one, an agent with $\lambda < 1$ will choose the boundary solution $\tilde{p}_0^{\,*} = 1$. The non-loss-averse agent of a high type cannot lower his beliefs. In the second part of the paper, I empirically test whether people adopt lower beliefs when expecting a signal, thus it is important to identify agents for whom the strict inequality does not hold.

3.1.6. Implications for the agent's bias. I define agent's *bias* as a difference between the chosen belief $\tilde{p}_0^{\,*}$ and the unmanipulated belief p_0. If p_0 fully captures the state of the world (agent's ability), this bias coincides with over- and underconfidence discussed extensively in the literature (see, e.g., Bénabou and Tirole, 2016). This implies $p_0 = 0$ if an agent has low ability, and $p_0 = 1$ for a high-ability individual. As explained in the previous section, a high-ability agent cannot boost his beliefs any further, hence he cannot be overconfident, whereas a low-ability individual cannot be underconfident. One can obtain the formula for agent's bias by moving p_0 to the other side of equation (3.2):

$$\tilde{p}_0^{\,*} - p_0 = \frac{1}{\gamma}\Big(1 - P(s=H|p_0)\,\eta - P(s=L|p_0)\,\lambda\eta\Big)(u_H - u_L). \tag{3.16}$$

Assuming $\eta = 1$, $\gamma > 0$, and $u_H > u_L$, the middle term is larger than zero for $\lambda < 1$ and lower than zero for $\lambda > 1$. The bias is positive for non-loss-averse and negative for loss-averse agents. Moreover, in our model, loss aversion (non-loss aversion) is a necessary condition for the negative (positive) bias to occur.

Prediction B.1

If p_0 is a degenerate belief that reflects the state of the world, the following is true:
i) an agent is overconfident if and only if he is non-loss-averse and low-ability,
ii) an agent is underconfident if and only if he is loss-averse and high-ability.

The model links the direction of agent's bias to his actual ability and the attitude towards losses in belief-based utility. Moreover, the agent's bias is varying with the probability of receiving a "good" or a "bad" signal. One can rewrite the formula for agent's bias using the

signal structure introduced in Section 3.1.3:

$$\tilde{p}_0^{\,*} - p_0 = \frac{1}{\gamma}(p_0 - 2cp_0 + c)(1-\lambda)(u_H - u_L), \tag{3.17}$$

where c is signal precision, $c > \frac{1}{2}$. For overconfident agents, the middle term reduces to $(p_0 - 2cp_0 + c) = c$. For underconfident individuals, this term becomes $(1-c)$. It is important to note that, since $c > \frac{1}{2}$, we have $c > 1-c$. The absolute bias, understood as the distance between the manipulated belief and the true state, will be larger for overconfident than for underconfident agents. Secondly, the absolute bias will be more responsive to changes in λ for overconfident agents. One can show this by comparing the derivatives taken with respect to λ, but it is also clearly visible in (3.16): lowering the probability of receiving a "bad" signal reduces the effect of λ on agent's bias. Since overconfident and underconfident agents are characterized by $p_0 = 0$ and $p_0 = 1$, and the signals are informative, the weight placed on λ is higher for overconfident agents. Taken together, this brings us to formulate the following prediction:

Prediction B.2

i) The absolute bias is larger for overconfident than for underconfident agents.
ii) The absolute bias is decreasing in the loss aversion parameter λ for overconfident agents, and increasing in λ for underconfident individuals.
iii) λ has a stronger impact on the absolute bias for overconfident agents.

Predictions B.1 and B.2 hold regardless of whether the agent expects to get a signal or not, because the agent's confidence type is not affected by treatment manipulation. We do not examine how the agent's bias changes in the two conditions – there is little to learn beyond what was presented in the previous sections. The above predictions are more interesting, as they demonstrate a link between overconfidence and loss aversion (B.1), and show how bias responds to changes in λ for different types of agents (B.2).

Another interesting implication of (3.17) is that the absolute bias is increasing in c for overconfident and decreasing in c for underconfident agents.[18] The precision of a signal affects agents differently, depending on their type. High-ability, underconfident agents become

[18]Our experiment was not designed to test this prediction, however, one could easily modify our design to check how the agent's bias changes when one increases the signal precision.

more accurate, whereas low-ability, overconfident individuals *less* accurate in their assessments.[19] For overconfident agents, the intuition is the following: if an agent is low-ability and the signal becomes more precise, the probability of receiving "good news" decreases, so the motive to lower one's beliefs to enhance utility from a positive surprise is reduced. The incentive to lower one's beliefs to avoid disappointment is still present but it is weighted with $\lambda < 1$, so the incentive to enjoy optimistic beliefs *right now* dominates. If I am almost sure that I will get a "bad" signal tomorrow, the best I can do is to enjoy the belief that I am smart today. An underconfident agent, on the other hand, expects to receive "bad news" with a lower probability, so a more precise signal decreases the incentive to adopt overly pessimistic beliefs, reducing the bias.

Although the set-up with two states of the world and $p_0 \in \{0, 1\}$ is very limited, one should note that the conditions described in this section are not a purely theoretical possibility. In any relative-performance setting, there are agents whose performance hits the upper or the lower bound. What can we say about those agents? The model suggests that agents with the lowest ability will be overconfident (on average, as those with $\lambda > 1$ will adopt $\tilde{p}_0^* = p_0 = 0$) and those with the lowest values of the loss-aversion parameter will be the most overconfident. At the same time, agents with the highest ability will be underconfident (on average), with their bias increasing in loss aversion.

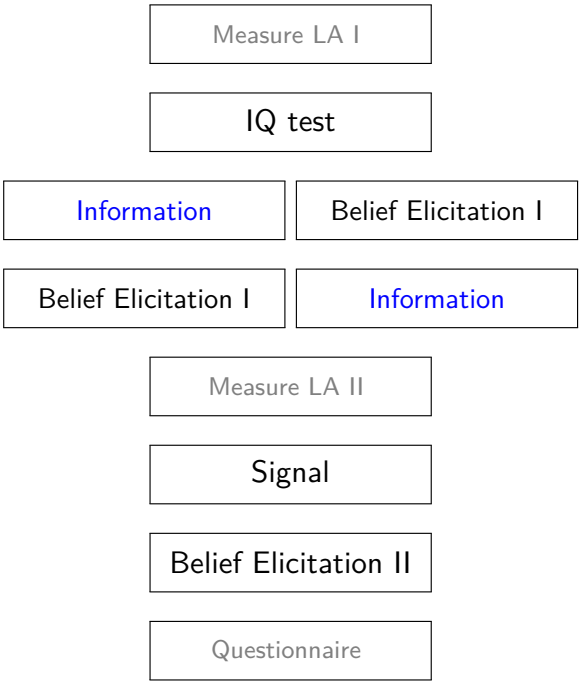
3.2. Experimental Design

In this section, I describe the experimental design and how it allows us to test the predictions formulated in Section 3.1. The outline of the experiment is presented in Figure 3.1 The experiment consists of several parts, which are the same in the Treatment and in the Control condition.[20] The two conditions differ only with respect to the timing of information provision (marked in blue in Figure 3.1). In the Treatment condition, the information about the signal was provided *before* the first belief elicitation, and in the Control condition, the information was given to subjects' *after* eliciting their beliefs.

[19] This might seem counter-intuitive. It is important to note that the bias we consider emerges *before* the information arrives as a result of a mechanism that is different from, e.g., asymmetric updating.

[20] Instructions for each part were distributed separately, and at the time of the IQ test and Questionnaire participants were not given any information about the remaining parts of the study. Before the IQ test, subjects were told that they will not receive the results of the test today, but this information will be available to them online, one week later. I followed the same procedure as Kozakiewicz (2020).

FIGURE 3.1. Outline of the experiment.



At the beginning of the study, participants solved an IQ test consisting of 29 standard logic questions. Participants were given 10 minutes to solve as many as they could knowing that they will be remunerated for their test score, which we will calculate based on the number of correctly answered questions minus the number of incorrect answers. Afterward, participants were either presented with a signal structure and information about an upcoming signal and then asked to report their beliefs about their relative performance (the Treatment condition), or were firstly asked to report their beliefs and only after they finished, they received the same information about the upcoming signal (the Control condition). The signal structure and elicitation procedure are described in detail in the following section.

Moreover, I introduced additional procedures to obtain two different measures of individual loss aversion. The first procedure was implemented at the very beginning of the experiment, before the IQ test (it is denoted with "Measure LA I" in Figure 3.1). The second procedure was implemented after the information provision and belief elicitation (denoted with "Measure LA II" in Figure 3.1). I describe these procedures in Sections 3.2.2 and

3.2.3. Once participants completed the stage “Measure LA II”, each subject received a signal as specified in the instructions. The signal was followed by the posterior belief elicitation (Belief Elicitation II)" and a two-page questionnaire.

3.2.1. Signal Structure and Belief Elicitation. I followed a procedure similar to the one developed in Kozakiewicz (2020). All participants were informed that in previous sessions over 300 subjects solved the same IQ test. Those participants were ranked according to their test scores and grouped into 10 groups which we refer to as “ranks”. Rank 1 was assigned to participants with the highest test scores, Rank 2 to participants with the second-highest test scores, and so on, up to Rank 10, which was assigned to subjects with the lowest scores.

In the Treatment condition, participants were instructed that, although their IQ test result will not be fully revealed to them during the session, they will receive a signal about their rank – the rank assigned to them by comparing their IQ test score to the scores of previous participants. I used the same signal structure as in Kozakiewicz (2020) and explained it in the same way. Participants were told that there are two boxes and each box contains 10 balls with numbers written on them. In the first box, the balls are numbered from 1 to 10 and each number occurs exactly once. In the second box, all balls have one number written on them, and this number is equal to the participant’s rank (the composition of the second box differs across subjects). For example, a person whose rank was Rank 4 will be facing two boxes presented in Figure 3.2. Subjects were informed that one box will be selected at random by the computer program (either box can be selected with equal probability) and one ball will be randomly drawn from the selected box. They will *not* get to know which box was selected

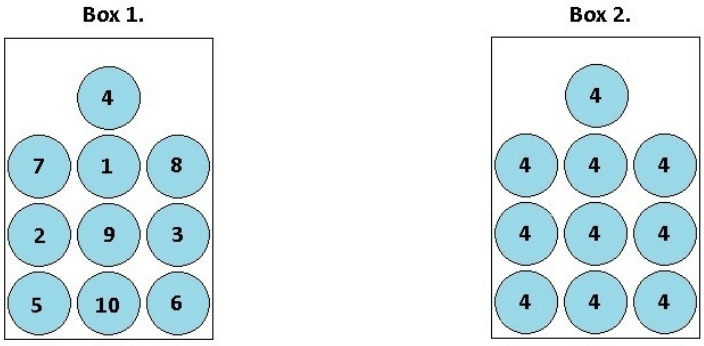


FIGURE 3.2. Boxes of a person whose rank was 4 (the “Signal” condition).

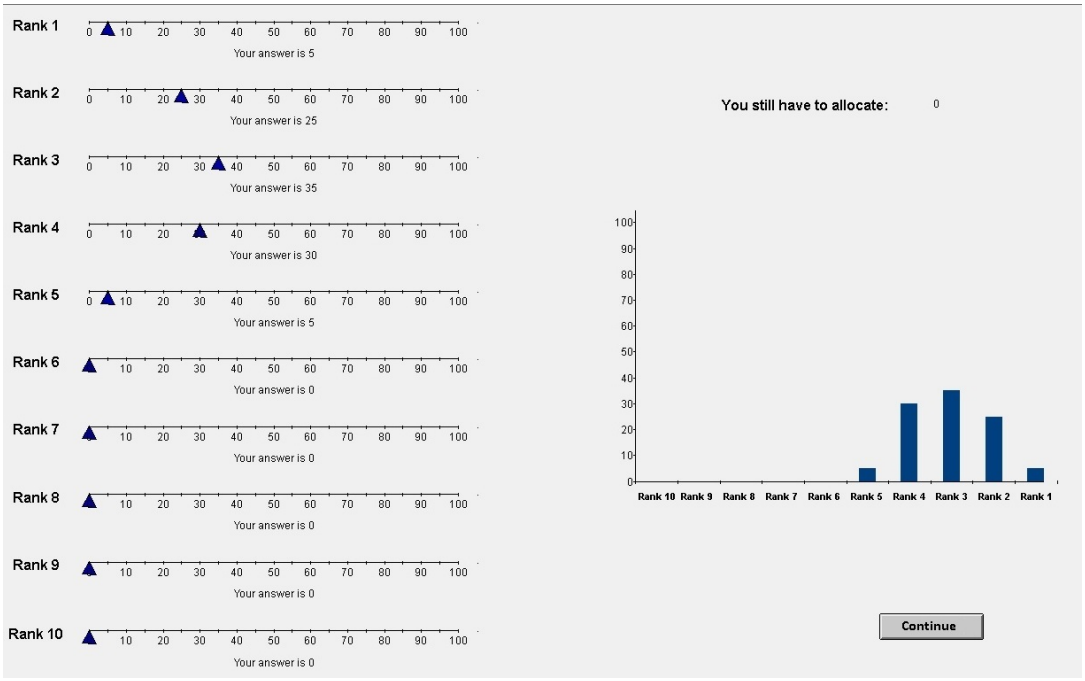


FIGURE 3.3. The screen-shot of the interface for belief elicitation.

by the program nor the composition of the boxes. Instead, the ball drawn for them will be displayed on their computer screens.

After providing information about the reference group and the signal, we explained to the subjects the belief elicitation task. Participants were asked to report their beliefs by allocating 100 points among the 10 ranks using a computer interface (see Figure 3.3). Subjects were allocating points by dragging blue arrows and were able to correct their choices as many times as they wished. The allocation immediately appeared on the graph to the right. The text above the graph informed participants how many points they still have to allocate before they can finish the task.

To incentivize truthful reporting, I used the Binarized Scoring Rule, following Hossain and Okui (2013). The formula was presented to the subject in a simple way, and we explained its implication: the chances of receiving a reward of 5 euros are maximized when one reports his true beliefs. Formally, the random variable X can take one of 10 values: (1,0,...,0,0), (0,1,...,0,0), ..., (0,0,...,1,0), (0,0,...,0,1); the position of 1 indicates the participant's rank. After receiving the agent's report $x = (x_1, ..., x_{10})$, where x_i denotes the share of points allocated to rank $i \in \{1, ..., 10\}$, I observe his true rank k, and the agent wins the prize if the QSR for multiple

events,

$$s(x,k) = 2x_k - \sum_i x_i^2 + 1,$$

exceeds a uniformly drawn random variable with the support $[0,2]$.

In the Control condition, participants were given the same information about the comparison group and were asked to report their beliefs about their rank. We use the same instructions for the belief elicitation as in the Treatment condition. The only difference between the two conditions is that in the Control condition, participants were not given information about the upcoming signal. The information about the upcoming signal, together with the explanation of the signal structure, was given to the subjects immediately after they finished the elicitation task.

3.2.2. Measuring Loss Aversion I. In order to obtain an independent measure of the loss aversion parameter λ_i, I designed a short survey that would capture the same concept. Before the IQ test, we presented subjects with a hypothetical, real-life scenario and asked them to answer several questions. The scenario that subjects were considering reads as follows:

Imagine that you took an important exam. After the exam, you are completely unsure whether the result will meet the expectations you have set for yourself. You think that there is a 50% chance that you will receive a score that you would consider a "bad" result (i.e. a result that does not meet your expectations) and a 50% chance that you will receive a score that you would consider a "good" result (i.e. a result that is equal to or better than your expectations). You will know the result in one week.

Today you will meet with someone who has more information about the outcome of your exam. This person cannot tell you the exact result but can give you a tendency whether you will get a "good" or a "bad" result. If the person tells you that you will receive a "good" result, it means that the probability that you will actually get a "good" result is 70% and the probability of getting a "bad" result is 30%. If the person tells you that you will get a "bad" result, it means that the probability that you will actually get a "bad" result is 70% and the probability of getting a "good" result is 30%. However, regardless of what the person will tell you, you cannot be absolutely sure about the result of the exam.

Participants did not receive paper instructions for this part – the hypothetical scenario and the questions appeared on their computer screens. They were asked to answer two questions intended to assess their utility before receiving a signal. The first question read: "How willing would you be to talk to that person about your score?" and the possible responses on a Likert scale ranged from 1="I would like to talk to this person about the test very much." to 9="I would not like to talk to this person about the test at all.". The second question read: "How would you feel right before that conversation?" and the responses ranged from 1="I would feel very relaxed before the conversation." to 9="I would feel very anxious before the conversation.".

I use the answers to the two questions as a measure of anticipatory utility that people experience before a signal. In light of our theory,

$$\begin{aligned} U_Q \;=\; & P(s=H|p_0)\,\left[u(p_1^H)+\eta\Big(u(p_1^H)-u(p_0)\Big)\right]+ \\ & +P(s=L|p_0)\,\left[u(p_1^L)+\eta\lambda\Big(u(p_1^L)-u(p_0)\Big)\right], \end{aligned} \tag{3.18}$$

where $u(p)=pu_H+(1-p)u_L$, as before. The equation is slightly different from (1). I assume that there is no need for belief manipulation when considering a hypothetical scenario. The agent takes the probabilities as given, and assesses his utility with $p_0=0.5$, $P(s=H|p_0)=0.5$, $p_1^H=0.7$, and $p_1^L=0.7$, with no manipulation costs. Furthermore, we asked about subjects' feelings over the upcoming signal, so there is no current belief-based utility component. After substituting the numbers, we obtain:

$$\begin{aligned} U_Q \;=\; & 0.5\left\{0.7u_H+0.3u_L+\eta\left[0.7u_H+0.3u_L-\Big(0.5u_H+0.5u_L\Big)\right]\right\}+ \\ & +\;0.5\left\{0.3u_H+0.7u_L+\eta\lambda\left[0.3u_H+0.7u_L-\Big(0.5u_H+0.5u_L\Big)\right]\right\}= \\ & =\;0.5\,(u_H+u_L)\;+\;0.1\,\eta\,(1-\lambda)\,\Big(u_H-u_L\Big). \end{aligned} \tag{3.19}$$

If we assume that u_H and u_L are the same for all participants, the differences in anticipatory utility reflect the differences in the loss aversion parameter λ. The two questions are intended

to capture these differences. Intuitively, a more loss-averse person would be 1) less willing to receive a signal, and 2) more anxious before the signal realization.[21]

3.2.3. Measuring Individual Loss Aversion II. I introduce two additional tasks to obtain another measure of individual loss aversion – one that could possibly validate the measure from the hypothetical scenario. The tasks were performed after Belief Elicitation I, but before subjects' received a signal. Recall that the utility in Period 0 (before seeing the signal realization) is:

$$U_0 = u(\tilde{p}_0) + P(s=H|p_0) \underbrace{\left[u(p_1^H) + \eta \left(u(p_1^H) - u(\tilde{p}_0) \right) \right]}_{U_g \text{ (utility after a "good" signal } s=H)} + \tag{3.20}$$
$$+ P(s=L|p_0) \underbrace{\left[u(p_1^L) + \eta\, \lambda \left(u(p_1^L) - u(\tilde{p}_0) \right) \right]}_{U_l \text{ (utility after a "bad" signal } s=L)} - \frac{\gamma}{2}(\tilde{p}_0 - p_0)^2,$$

where $\tilde{p}_0$ is the manipulated belief elicited in the first task. If one knew U_g and U_l, as well as p_1^H and p_1^L (subjects' posterior beliefs after "good" and "bad" signals), one could back out the loss aversion parameter λ. However, the parameter derived in this way would be a function of the manipulated belief $\tilde{p}_0$.

In order to obtain a measure of λ_i that does not depend on the manipulated belief, I use incentivized, conditional choices.[22] I designed two additional tasks to measure the monetary equivalent of U_g and U_l, the posterior beliefs p_1^H and p_1^L, as well as the unmanipulated prior p_0 conditioning on a signal realization.

Posterior Beliefs. To elicit prior and posterior beliefs using conditional choices, I follow Kozakiewicz (2020) and ask participants to report their beliefs about the source of a signal conditioning on signal realization. For every $x \in \{1,...,10\}$, subjects were asked to consider each number *as if* this number was actually drawn and report their subjective probability that the number came from Box 2. Subjects made their reports by allocating 100 points between Box 1 and Box 2 using a computer interface (see Figure 3.4) and were incentivized to report truthfully with binarized scoring rule (Hossain and Okui, 2013). While participants

[21] I assume that there is no heterogeneity in how agents value partial information beyond anticipatory utility. If people derive utility from being more certain about *any* outcome, this utility is the same for all agents. The differences in preferences over the degree of uncertainty are not captured by our model.

[22] I assume that there is no need for belief manipulation when making a decision *conditioning* on the signal realization. This assumption is similar to the assumption considered in Section 3.2.2: there is no need for belief manipulation when considering a hypothetical scenario. Both assumptions are based on the idea that belief manipulation emerges in a state of emotional arousal (Kozakiewicz, 2020).

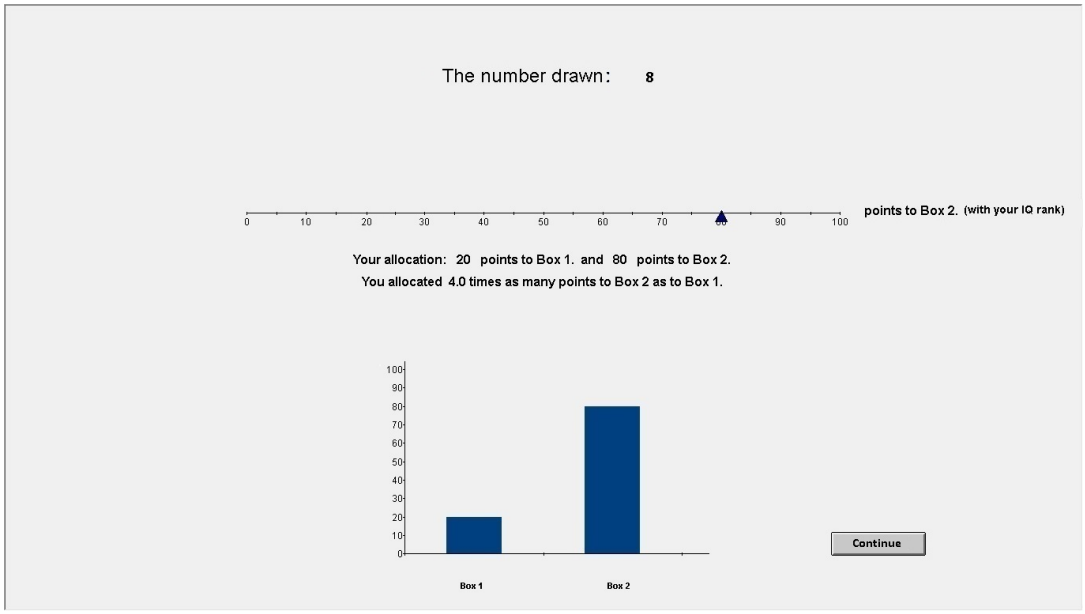


FIGURE 3.4. The screen-shot of the interface used in hypothetical questions.

were free to choose any allocation, we instructed them on how to arrive at the Bayesian posterior given one's prior belief distribution. Moreover, we explained that the choices they are making are not entirely hypothetical. At the end of the session, one box will be selected by the computer program and one ball will be randomly drawn from the selected box. Although they will not see the number drawn, their payoff will depend on the decision they made for this number. For the details of the procedure, see Kozakiewicz (2020).

Monetary Equivalents. Secondly, we elicited subjects' willingness to pay to avoid a signal x. For every $x \in \{1,...,10\}$, participants were presented with a price list, using which they decided between two options. Option A was: "Do not see the number x and get y Euro", whereas Option B was: "See the number x and get 2 Euro", with y ranging from 1 to 3 in increments of 10 cents. Participants were informed that this task will be implemented with probability 50% and with probability 50% they will be given the signal as specified earlier in the instructions. If this task is implemented, then one signal will be drawn at random and the decisions they made for this signal will apply. Specifically, one of the 21 lines in the price list will be drawn at random, and the decision they made in this line will be realized. Participants' decisions provide a measure of U_g and U_l – the utilities they experience after a "good" and a "bad" signal. In Appendix D, I describe how to use subjects' responses to identify λ_i.

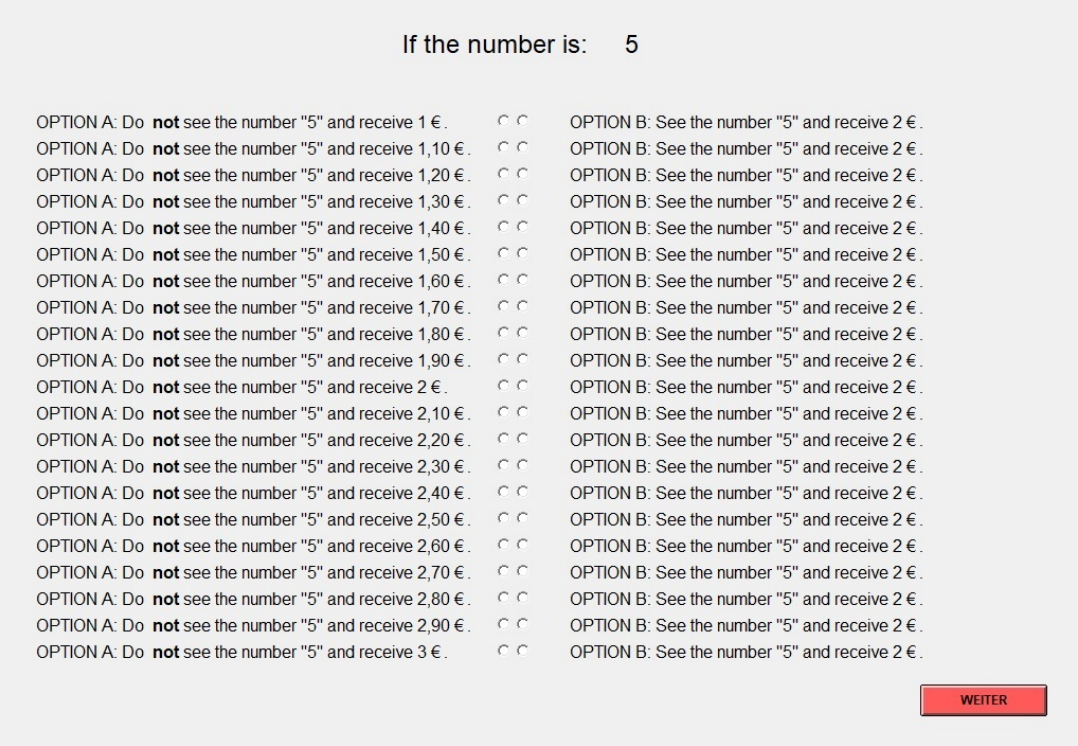


FIGURE 3.5. The screen-shot of the interface used by subjects in WTP.

3.3. Testable Predictions

Belief Elicitation I provides us with our main outcome variable: the optimal prior $\tilde{p}_0^*$. By comparing the average belief in the Treatment and in the Control condition, we shed light on Prediction 1. One would expect the average prior in the Treatment condition to be lower (more pessimistic) than the average prior in the Control condition.

Hypothesis 1

The average prior belief chosen in the Treatment condition is lower (more pessimistic) than the average prior belief chosen in the Control condition.

The model predicts that there should be a negative correlation between the prior belief $\tilde{p}_0^*$ and the measure of loss aversion λ in the Treatment condition. More loss-averse subjects should adopt lower beliefs compared to less loss-averse subjects (Prediction 2.1). In the Control condition, this correlation should be weakly negative due to the discount factor $\epsilon \in (0,1)$. Moreover, the negative effect should be stronger in the Treatment condition (Prediction 2.2).

In order to test these predictions, we look at the following regression:

$$Prior_i = \alpha_0 + \alpha_1\, Treatment_i + \alpha_2\, \lambda_i + \alpha_3\, Treatment_i \times \lambda_i + \epsilon_i, \tag{3.21}$$

where $Prior_i$ denotes subject i's reported prior belief, $Treatment_i$ is an indicator variable taking value 1 if participant i was in the Treatment condition and 0 otherwise, λ_i is a measure of participant i's loss aversion. The coefficient α_2 informs us about the correlation between subjects' prior beliefs and their gain-loss attitudes in the Control condition ($Treatment_i = 0$). It is expected to be negative or equal to zero. The correlation between beliefs and gain-loss attitudes in the Treatment condition is captured by $\alpha_2 + \alpha_3$. Based on Prediction 2.2, we expect this sum to be negative. The coefficient α_3 at the interaction term informs us whether an increase in the loss aversion parameter has a larger negative effect in the Treatment compared to the Control condition. We expect this coefficient to be negative.

Hypothesis 2

i) In the Treatment condition, there is a negative correlation between subjects' loss aversion and their prior beliefs. Coefficient $\alpha_2 + \alpha_3$ in (3.21) *is negative.*

ii) In the Control condition, the correlation is weakly negative. The coefficient α_2 in (3.21) *is negative or no different than zero.*

iii) The loss aversion parameter has a larger negative effect in the Treatment condition compared to the Control condition. The coefficient α_3 in (3.21) *is negative.*

In order to test Prediction 4.1, I use a specification that allows for examining a joint effect of the loss aversion parameter λ and unmanipulated beliefs p_0. Note that we did not elicit unmanipulated beliefs in the experiment. However, it is reasonable to assume that these beliefs are correlated with agents' actual cognitive ability. Therefore, I use subjects' rank, which is a measure of one's cognitive ability, as a proxy for the former. I consider the following regression:

$$Prior_i = \beta_0 + \beta_1\, Ability_i + \beta_2\, \lambda_i + \beta_3\, Ability_i \times \lambda_i + \epsilon_i, \tag{3.22}$$

where λ_i is the measure of participant i's loss aversion. The coefficient β_3 captures the joint effect of λ and subject's ability on prior beliefs. According to Prediction 4.1, this effect should

be positive. Considering that an increase in λ has a negative effect on prior beliefs, an increase in participant's ability counteracts this force – it lessens the negative effect of loss aversion.

Hypothesis 3

The effect of loss aversion on prior beliefs is less negative for subjects with higher ability. The coefficient β_3 in (3.22) *is positive.*

Next, I examine the joint effect of the treatment and 1) the loss aversion parameter, 2) unmanipulated beliefs. I construct the following regression:

$$\begin{aligned} Prior_i = {} & \gamma_0 + \gamma_1\, Treatment_i + \gamma_2\, \lambda_i + \gamma_3\, Ability_i + \\ & + \gamma_4\, Ability_i \times \lambda_i + \gamma_5\, Treatment_i \times \lambda_i + \gamma_6\, Treatment_i \times Ability_i + \\ & + \gamma_7\, Treatment_i \times Ability_i \times \lambda_i + \epsilon_i. \end{aligned} \tag{3.23}$$

In line with Prediction 4.2, an increase in the unmanipulated belief p_0 weakens the negative effect of λ on prior beliefs to a larger extent in the Treatment condition. This relation is captured by the coefficient γ_7 in (3.23). A positive coefficient implies that the mitigating effect is stronger in the Treatment than in the Control condition.

Hypothesis 4

An increase in the unmanipulated belief p_0 lessens the negative effect of the loss aversion parameter on the optimal belief to a larger extent in the Treatment than in the Control condition. The coefficient γ_7 in (3.23) *has a positive sign.*

Since our measure of λ is ordinal, not cardinal, I cannot use it to test Predictions 3.1 and 3.2. In the second part of the analysis, I divide subjects into two groups: loss-averse and non-loss-averse. A subject is classified as loss-averse if the average of his responses to the two hypothetical questions was above the neutral value of 5. In the view of the fact that this classification is discretionary, I delegate the tests I perform based on it to Appendix A. In the main body of the paper, I describe tests performed using a discrete variable: subjects' responses on

the Likert scale. I use the categorical variable only to control for the boundary cases, that is, participants for whom Predictions 1-4 do not necessarily hold, as explained in Section 3.1.5.

3.4. Data Analysis

In this section, I present the results of the data analysis. First, I briefly describe raw data. In Sections 3.4.2 and 3.4.3.1, I test Hypotheses 1-4 using two measures of loss aversion. In what follows, I restrict the sample to the participants who correctly answered at least two out of four control questions. Out of 234 participants, 7 subjects made mistakes in three or four questions (they constitute 3% of the sample). The results remain very similar, albeit noisier, if I include those subjects in the analysis (see Appendix B).

3.4.1. Raw Data. The distribution of participants' IQ test scores is presented in Figure 3.6. The average test score was 5.17, with a standard deviation of 3.42. Table 3.1 presents the average score and the average rank of participants in the Treatment and the Control condition. There is a difference in the average IQ test score in the two conditions, which translates to a difference in assigned ranks (see Table 3.1).[23]

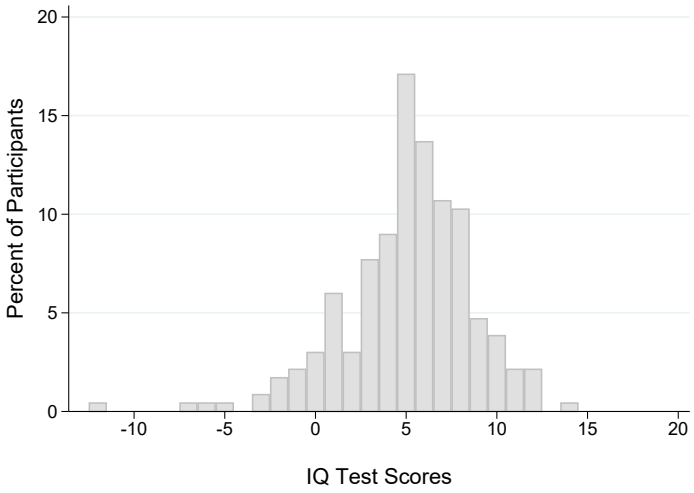


FIGURE 3.6. Distribution of the IQ test scores.

[23]Due to the differences in the instructions, we could not randomly assign the treatment status to participants *within* a session. We randomized at the session level, alternating between the Treatment and the Control condition. The small number of sessions is likely to be the reason behind the differences in the IQ test scores. There is no significant difference in other measures we collected: loss aversion, cognitive reappraisal, and suppression.

TABLE 3.1. Differences between participants in the two conditions.

	Treatment	**Control**		Diff < 0	Diff ≠ 0	Diff > 0
IQ test score	4.81	5.51	*p-value:*	0.940	0.121	0.061
Rank	5.99	5.34	*p-value:*	0.964	0.073	0.036
Loss Aversion	4.93	4.75	*p-value:*	0.778	0.443	0.222
N	110	117				

The average mean of belief distributions reported by participants in Belief Elicitation I is 4.66, while the average median equals 4.65 (both values expressed in terms of individual rank). However, only 14 participants revealed a symmetric belief distribution, the remaining subjects reported right- or left-skewed distributions.[24] The average range was 5.1, meaning that, on average, participants' belief distributions span over 5 ranks. There is a positive and significant correlation between subjects' beliefs and their true rank (the Pearson correlation coefficient $r = 0.287$ is significant at the 1%-level).

Participants in our sample tend to be *overconfident*, revealing belief distributions with means that were, on average, 1 lower than their true rank.[25] Note that, in our set-up, a lower rank corresponds to a better performance. Similarly, a lower mean belief (expressed in terms of rank) indicates a belief in a higher performance. To avoid confusion, I reverse the two variables so that they are increasing in the agent's ability and perceived ability. The new variable takes values from 1 to 10 and denotes the decile of the IQ test score distribution that the agent's score fell into (or was believed to fall into), with higher values corresponding to a better performance.

Our first measure of loss aversion is defined as the average response to the two hypothetical questions. Its distribution is presented in Figure 3.7. The distribution has an inverted U-shape with a mean of 4.84 and a standard deviation of 1.76. There is no significant difference in the average loss aversion between the Treatment and the Control condition (see

[24] Almost as many participants revealed a right-skewed distribution as a left-skewed distribution. The average difference between the mean and the median in these groups was 0.27 and -0.24, respectively.

[25] I define a participant to be overconfident if his mean belief was lower than the actual rank (lower rank corresponds to a better performance). Using this definition, I classified 60% of participants as overconfident and 40% as underconfident. There was only one subject whose mean belief exactly matched the actual rank (assigning him to either group does not change the results presented later in the paper). The distribution of subjects' bias can be found in Appendix E.

FIGURE 3.7. Distribution of loss aversion in the sample.

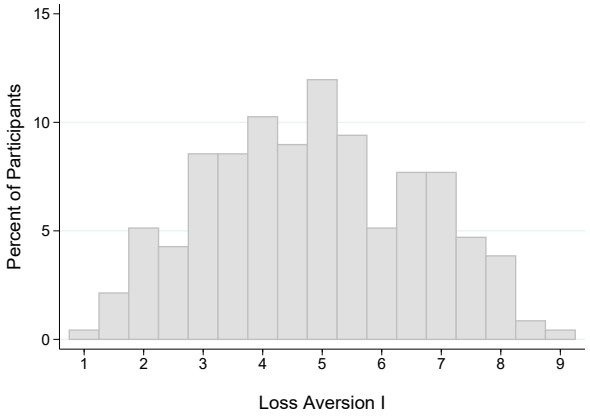


Table 3.1). At the same time, our measure of loss aversion correlates with beliefs as predicted by the model – I explore this relationship in the following sections.

3.4.2. Testing Model Predictions. In this section, I test the model predictions using the experimental data. As explained in Section 3.1.5, the model predictions do not hold for 1) loss-averse agents with the lowest ability, and 2) non-loss-averse agents with the highest ability. For this reason, I always present two sets of results. The first results are based on the sample of all participants. The second set is based on a restricted sample, which is created by excluding 1) subjects with the lowest rank who were classified as loss-averse, and 2) subjects with the highest rank classified as non-loss-averse. I classify a subject as loss-averse if his average response to the two hypothetical questions was above 5 – a value that is the median of distribution presented in Figure 3.7 and, at the same time, the mid-point on a 9-point Likert scale that we used. The remaining subjects were classified as non-loss-averse.[26]

3.4.2.1. *Hypothesis 1.* First of all, I examine the effect of being assigned to the Treatment condition on prior beliefs about cognitive ability. The dependent variable is the mean of individual belief distribution reported in Belief Elicitation I.[27] It takes values from 1 to 10, and a higher value indicates a belief that one obtained a higher test result.[28] I regress the

[26]In Appendix B.4, I present the results based on alternative definitions: 1) a subject is classified as loss-averse if his response to each of the two questions was above the median response, and 2) 40% of subjects with the lowest loss aversion are classified as "non-loss-averse" and the rest as "loss-averse" (the ratio is set to match the results in Goette et al., 2019). Our results are robust to these changes.

[27]The results are the same if I use the median, the 1^{st} or the 3^{rd} quartile instead (see Appendix B).

[28]For example, a subject who revealed the mean belief of 9 believes that he was among the best performers – his IQ test score was better or equal to the IQ test scores of 90% other participants.

TABLE 3.2. The effect of treatment on mean beliefs.

	All subjects	BC	R (All - BC)
Treatment	-0.141	0.549	-0.249
	(0.225)	(0.745)	(0.233)
Const.	6.408***	5.793***	6.506***
	(0.157)	(0.509)	(0.163)
N	227	30	197

The dependent variable is the mean belief revealed in Belief Elicitation I. "All subjects" refers to the whole sample. "BC" are the boundary cases – subjects who do not fulfill the conditions necessary for the negative effect to occur (see Section 3.1.5). "R": the remaining subjects (all subjects minus boundary cases). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

dependent variable on a treatment dummy. The first column contains the estimates based on the sample of all participants. The coefficient at the "Treatment" variable has the predicted sign but is not significant.

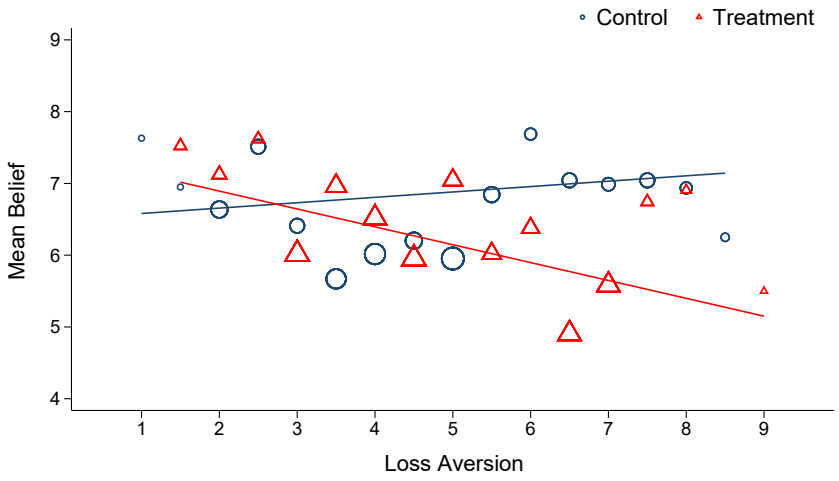
However, it is important to note that the model does not predict a negative treatment effect for all subjects. As explained in Section 3.1.5, the treatment manipulation should not affect 1) loss-averse participants with the worst performance, and 2) non-loss-averse subjects with the best performance. In the second column in Table 3.2, I show that, indeed, the direction of the effect for those subjects is inconsistent with Prediction 1. Therefore, I restrict the sample to the subjects who fulfill the assumptions of the model necessary for the effect to occur. In the last column in Table 3.2, I present the estimates based on the restricted sample. The treatment effect is negative, as predicted by the theory, but not significant (p-value of one-sided t-test = 0.287). More data is needed to confirm Hypothesis 1.

Result 1

The average prior belief reported in the Treatment condition is not significantly lower than the average prior belief in the Control condition.

3.4.2.2. *Hypothesis 2.* Before I test Hypothesis 2, I plot the averages of subjects' beliefs for groups with different loss aversion parameters (see Figure 3.8). Observations in the Treatment condition are denoted with red triangles and in the Control condition with navy circles. Their sizes correspond to the relative frequencies. The blue and red line corresponds to the regression fitted on the data from the Treatment and the Control condition, respectively. One can notice that, for more loss-averse subjects, the average beliefs in Treatment tend to be below the averages in Control (the triangles are below the circles on the right side of the graph). For subjects with lower values of the loss aversion parameter, the averages in Treatment are slightly above or equal to the averages in Control.

FIGURE 3.8. The mean belief and loss aversion in the two conditions.



However, the graph does not take into account the differences in ability between the two groups. For this reason, I turn to the regression analysis. The results are presented in Table 3.3. The first two columns contain the estimates based on the entire sample, and the last columns contain the estimates based on the restricted sample. The dependent variable is the mean of individual belief distribution revealed in Belief Elicitation I. In the first specification, I regress it on an indicator variable "Treatment", a discrete variable "Loss Aversion" (which takes values between 1 and 9 in a step of 0.5), and their interaction. In the second specification, I add a control for individual rank.

TABLE 3.3. The effect of treatment and loss aversion on mean beliefs.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	All subjects		Restricted sample (R)	
	(1)	(2)	(1)	(2)
Treatment	1.038	1.049*	1.240*	1.114*
	(0.657)	(0.632)	(0.664)	(0.648)
Loss Aversion	-0.021	-0.022	0.075	0.020
	(0.085)	(0.082)	(0.089)	(0.088)
Treatment × Loss Aversion	-0.238*	-0.218*	-0.311**	-0.271**
	(0.127)	(0.123)	(0.131)	(0.128)
Ability		0.173***		0.165***
		(0.040)		(0.049)
N	227	227	197	197

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Ability" takes values from 1 to 10 and denotes the position in the IQ test distribution (with 10 assigned to subjects with the highest test scores). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

First, the sum of coefficients at the "Loss Aversion" variable and its interaction informs us about the correlation between the mean belief and the loss aversion parameter in the Treatment condition. I reject the null hypothesis that the sum of coefficients is larger or equal to zero (p-value of the one-sided t-test = 0.004). Second, the coefficient at the "Loss Aversion" variable in Table 3.3 is not significantly different from zero. There is no correlation between the mean belief and the loss aversion parameter in the Control condition. Both results are in line with the model predictions. Lastly, the coefficient at the interaction term is negative and significant, as predicted by the model. An increase in the loss aversion parameter by 1 point translates to a 0.218 decrease in the mean belief reported in the Treatment compared to the Control condition. In relative terms, this effect corresponds to 5% of the average belief in the sample or 13% of its standard deviation. The effect is stronger when I restrict the sample as prescribed by the theory. With the p-value of 0.036, the interaction effect is significant at the 5% level. Based on these results, I confirm the second prediction of the model

(Hypothesis 2).[29]

Result 2

i) In the Treatment condition, there is a negative correlation between the loss aversion parameter and prior beliefs. As predicted, more loss-averse subjects adopt lower beliefs.
ii) In the Control condition, there is no correlation between the two variables.
iii) The interaction effect is negative and significant, in line with the model predictions. More loss-averse subjects lower their beliefs more in the Treatment compared to the Control condition.

3.4.2.3. *Hypothesis 3 and 4.* In order to test Hypothesis 3, I run the regression specified in (3.22). The results are presented in Table 3.4. The dependent variable is the mean of individual belief distribution. In the first specification, I regress it on subjects' ability, loss aversion, and their interaction. "Ability" and "Loss Aversion" are defined in the same way as before. Additionally, I control for whether an individual was assigned to the Treatment condition (the results are the same without this control). The coefficient at the interaction term is positive and significant, which is consistent with the model prediction: an increase in the unmanipulated beliefs (proxied by ability) lessens the negative effect of the loss aversion parameter on the optimal belief. The result prevails if we add the remaining two-way interactions (it is worth noting that the coefficient at "Treatment × LA" in Specification (2) is negative and significant as in Table 3.3, providing support for Result 2).

Result 3

The effect of loss aversion on the mean belief is less negative for participants with higher ability. The coefficient β_3 in (3.22) *is positive.*

Lastly, I consider the triple interactions with the "Treatment" variable. An increase in ability should reduce the negative effect of loss aversion to a larger extent in the Treatment compared to the Control condition. The coefficient at the triple interaction "Treatment × Ability × Loss Aversion" should be positive and significant. As we see in Table 3.4, the coefficient is

[29] One might wonder whether the results could be driven by a small number of participants with extremely inaccurate beliefs about themselves. In Appendix B.3, I show that this is not the case.

TABLE 3.4. The effect of treatment, loss aversion, and unmanipulated beliefs.

	All subjects			Restricted sample (R)		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	0.011	0.507	-0.445	-0.111	0.461	0.480
	(0.215)	(0.767)	(1.437)	(0.227)	(0.773)	(1.622)
Loss Aversion	-0.403***	-0.297**	-0.377**	-0.477***	-0.325*	-0.323
	(0.136)	(0.150)	(0.181)	(0.173)	(0.185)	(0.238)
Ability	-0.083	-0.107	-0.178	-0.126	-0.187	-0.186
	(0.118)	(0.123)	(0.152)	(0.140)	(0.147)	(0.190)
Ability × LA	0.053**	0.050**	0.064**	0.065**	0.063**	0.063
	(0.023)	(0.023)	(0.029)	(0.028)	(0.028)	(0.038)
Treatment × LA		-0.194	0.000		-0.279**	-0.283
		(0.122)	(0.275)		(0.128)	(0.343)
Treatment × Ability		0.083	0.262		0.139	0.136
		(0.079)	(0.242)		(0.098)	(0.278)
Treatment × Ability × LA			-0.037			0.001
			(0.047)			(0.057)
N	227	227	227	197	197	197

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the hypothetical questions.
"Ability" is subject's position in the test score distribution taking values from 1 to 10, with 10 assigned to participants with the highest test scores. All specifications include a constant (omitted for clarity).
"All" refers to the sample of all participants, and with "R" I denote the restricted sample.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

not statistically significant. Based on these results, we cannot confirm Hypothesis 4 in our sample.

Result 4

Contrary to the model prediction, an increase in ability does not reduce the negative effect of loss aversion to a larger extent in the Treatment compared to the Control.

There might be several reasons why the last effect is not present in our dataset. First, one can reckon that the treatment manipulation is rather subtle. Detecting an effect might require a larger sample size, especially since the comparisons are made between subgroups. Further data collection, preceded by a power analysis, would be necessary.

Second, while we take into account that participants with the highest or the lowest rank are constrained in their manipulation, the same is true (to some extent) for subjects with the second-highest or the second-lowest rank and so on. This likely contributes to the noisiness of the data and cannot be corrected using standard methods for analyzing censored data. People for whom the strategy of lowering beliefs is unavailable (probably not only in the experiment but throughout life) are likely to develop different ways of dealing with unpleasant news. My recommendation would be to collect enough data to be able to perform the analysis on a sample of people in the middle of the distribution, and separately examine subjects with the highest and the lowest ability.

3.4.3. Measure of Loss Aversion II. In this section, I report the results of eliciting the willingness to pay for signals. Unfortunately, the attempt to retrieve the loss aversion parameter from subjects' choices was not entirely successful.[30] Participants in our experiment were not willing to forgo money – even as little as 10 cents – to increase the probability of receiving (or not receiving) a signal. Depending on the signal realization, between 76% and 82.5% of all participants followed a strategy of maximizing monetary outcome by choosing Option B in the first 10 or 11 lines, and then switching to Option A.[31] In the line number 11, subjects were deciding between "Option A: Do not see the number x and receive 2 euro." and "Option B: See the number x and receive 2 euro.". There is some variation in whether subjects chose Option A or B in this line, which I exploit in the analysis below.[32]

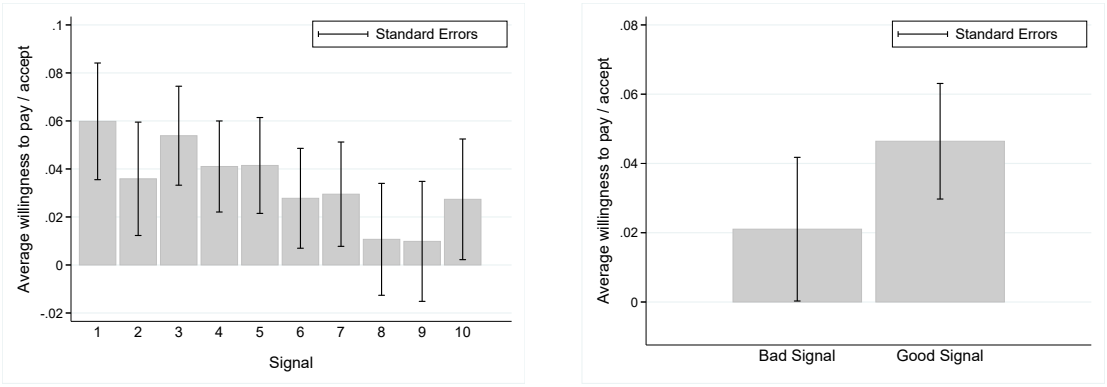
In Figure 3.9, I present the average willingness to pay for a signal depending on its realization. Recall that the price lists were designed with 10-cent steps and participants could pay or accept up to 1 euro (starting with an endowment of 2 euro). I assume the switching point to be the mid-point between the monetary gain or loss relative to the endowment in the last line in which Option B was chosen, and the monetary gain or loss in the first line in which

[30] I describe in detail how one can retrieve loss aversion parameters from choice data in Appendix D.

[31] 65% of subjects followed the profit-maximizing strategy in *every* decision they made.

[32] Suggestions for improvement include 1) changing the design of the price lists, and 2) increasing the relevance of a signal. As for 1), I would suggest using a price list that requires less effort, for example, reducing the number of lines by using the staircase method or asking participants to directly enter their price. Each of these options has its disadvantages, however, I think that reducing the effort required to fill in multiple price lists is crucial to improve the measurement. Secondly, one could think about making the signal more relevant. Increasing the precision of a signal is one possibility, however, one should keep in mind that very precise signals might have the opposite effect: subjects might be willing to acquire them regardless of their loss aversion to shorten the painful waiting period. Sweeny and Falkenstein (2015) show that, in some cases, waiting for an adverse event is more unpleasant than experiencing it. Another possibility is postponing revealing test results indefinitely, making the signal the only opportunity to receive information about the result.

FIGURE 3.9. The average willingness to pay for a signal.



(A) Signals 1-10

(B) "Good" vs "Bad" Signals

Option A was chosen.[33] For example, if a participant chose Option B in every line up to 1.70 euro, and chose Option A for the first time when given a choice between "Option A: Do not see the number x and receive 1.80 euro" and "Option B: See the number x and receive 2 euro", I assume that his willingness to pay/accept is $\delta_x = -0.25$ euro. The negative (positive) sign indicates that he is willing to pay (accept) 25 cents for not seeing the number. In other words, a negative δ indicates that the subject is willing to pay money to *decrease* the probability of seeing the number, while positive values indicate that he is willing to forgo money to *increase* the probability of seeing the number.

In Panel (A) in Figure 3.9, we see that, on average, subjects were willing to forgo 6 cents to see a signal "1", while the amount is not significantly different from zero for signals "8" or "9".[34] The difference is statistically significant and consistent with our theory – high signals are more desirable as they entail higher belief-based utility. In Panel (B) in Figure 3.9, I present the average willingness to pay for signals below or equal to 5 ("Good Signal") and signals above 5 ("Bad Signal"). With the p-value of 0.119, the difference is not significant at any acceptable level. Note that this definition does not take into account subjects' prior beliefs about their ability. A signal "5" might be considered a "good" signal by someone who believes

[33] If a subject chose Option A (Option B) in every decision, he is assumed to have the willingness to pay/accept $\delta = -1.05$ ($\delta = 1.05$).

[34] According to the definition, a person who prefers to keep 2 euro and decides to see the number is assigned the willingness to pay of 5 cents. In other words, the value of 0.05 (-0.05) was assigned to a subject who maximized the monetary gain and only chose to see (not see) the signal at the point of monetary indifference.

to rank above 5, and a "bad" signal for a person believing to be in the top deciles. When I define signal valence individually, based on one's beliefs revealed in the hypothetical choices, the difference grows larger and becomes significant (p-value of a one-sided t-test = 0.054). Consequently, defining the signal valence in absolute terms might lead to misclassification. In the following section, I choose a middle ground and define a "bad" signal to be one of the three worst signals ("8", "9", or "10"). This definition allows us to have the same number of "bad" signals for each participant, while only 3 subjects could be potentially misclassified. As a robustness check, I conduct the same analysis using the other two definitions (signal valence based on individual beliefs or five worst signals) and present the results in Appendix C.

3.4.3.1. *Results Based on Loss Aversion II.* Although I cannot derive a precise measure of loss aversion using subjects' willingness to acquire signals, I can coarsely classify participants as avoiding negative information (possibly due to loss aversion) based on their decisions in the price lists. Afterward, I test the model predictions using this classification. I define a person as avoiding information if she was willing to forgo any amount of money to *not receive* a "bad" signal or, in case of monetary indifference, decided to *not see* a signal. I look at the willingness to pay for signals "8", "9", and "10", that is, the three worst signals.[35] I define a person as "loss-averse" if she avoided at least two out of the three worst signals.[36] There were 53 participants of this type; they constitute 30% of the restricted sample of 178 participants.[37] The measure of loss aversion described above coincides with the loss aversion based on the hypothetical questions for 60% of subjects.

I conduct the same analysis as in Section 3.4.2. The most important results are gathered in Table 3.5 (complete analysis can be found in Appendix C). First, the difference in average beliefs between the Treatment and the Control condition is negative but not significant. Second, I run the same regression as Specification (2) in Table 3.3 using the second measure of loss aversion. The coefficient at the interaction term is negative, as predicted by the model,

[35] In Appendix C, I conduct the same analysis for 1) the five worst signals (signals from 6 to 10), and 2) signals that were higher or equal to one's mean belief revealed through the hypothetical choices. The results are consistent with the estimates presented in this section.

[36] As a robustness check, I conduct the same analysis using a different definition: a person was classified as loss-averse if he or she avoided all three signals. This condition leaves us with 50 loss-averse subjects (23% of the restricted sample). The results are almost the same as presented in this section.

[37] Note that our model cannot explain a decision to pay any amount of money to *receive* a "bad" signal. I restrict the sample to subjects who did not contradict the theory (allowing for one mistake). Participants that violated this condition constituted 11.5% of the original sample. I present the results with and without this restriction in Appendix C. Moreover, as in the previous section, I exclude participants for whom the predictions of the model do not hold.

TABLE 3.5. Model predictions based on the second measure of loss aversion.

Hypothesis 1		
Coeff. at "Treatment" *(p-value)*	-0.191	*(0.437)*
Hypothesis 2		
Coeff. at "Treatment × Loss Aversion" *(p-value)*	-0.857	*(0.102)*
Hypothesis 3		
Coeff. at "Ability × Loss Aversion" *(p-value)*	0.140	*(0.205)*
Hypothesis 4		
Coeff. at the triple interaction *(p-value)*	0.284	*(0.202)*

and close to significant at the 10% level (p-value of one-sided t-test = 0.102). The coefficient is around 25% lower than the estimate based on the first measure.[38] I also use Specification (1) from Table 3.4 to examine the joint effect of loss aversion and unmanipulated beliefs. The coefficient at the interaction term is positive but not significant (p-value of one-sided t-test = 0.205). Although the results are much noisier than those based on the first measure, it is reassuring to see that the effects go in the same direction. Lastly, I re-run the regression from Specification (3) in Table 3.4. The coefficient at the triple interaction is positive but not significant (p-value of one-sided t-test = 0.202). More data of a better quality is needed to reconcile these findings.

3.4.4. Over- and Underconfidence. In this section, I examine the link between the measure of loss aversion and the agent's bias: I test the two predictions formulated in Section 3.1.6.

3.4.4.1. *Prediction B.1.* In order to test the first prediction, I classify subjects according to the relevant characteristics. I define a participant as "high-ability" if his rank is below or equal to 5 (lower ranks correspond to better results). Participants whose ranks were above 5 are classified as "low-ability". Furthermore, I follow my baseline definition of loss aversion. A subject is classified as "loss-averse" if the average of his responses to the hypothetical questions was above 5, which is the middle value on the Likert scale that I used. I look at

[38] I compare this coefficient to the results based on a binary variable gathered in Appendix A (in the main text, the "Loss Aversion" variable takes discrete values based on the 9-point Likert scale).

TABLE 3.6. Confidence, ability, and loss aversion.

	Overconfident	**Underconfident**
Non-Loss-Averse & Low-Ability	62 (94%)	4 (6%)
Loss-Averse & High-Ability	11 (28%)	28 (72%)

This table shows how many of non-loss-averse and low-ability subjects (loss-averse and high-ability subjects) are overconfident or underconfident.

two groups: 1) subjects who are "low-ability" and "non-loss-averse", and 2) subjects who are "high-ability" and "loss-averse". Participants in these groups constitute 50% of the sample. The remaining 50% include subjects who are: 3) low-ability and loss-averse, 4) high-ability and non-loss-averse. Due to the presence of 3) and 4), the strong version of Prediction B.2 does not hold. Still, I find it worthwhile to test a weaker version of our hypothesis: if an agent is low-ability and non-loss-averse (high-ability and loss-averse) then he is *more likely* to be overconfident (underconfident), as well as its converse. In Table 3.6, I show how many low-ability and non-loss-averse (high-ability and loss-averse) subjects were overconfident (underconfident). Table 3.7 shows what fraction of overconfident subjects belong to group 1) versus the remaining groups. Although the numbers seem to be in line with our hypotheses – 94% of subjects in group 1) tend to be overconfident, and 72% of subjects in group 2) tend to be underconfident – we cannot draw conclusions based on these fractions alone. Because I am using relative performance, which is bounded, some relations will arise mechanically. In particular, low-ability individuals will be more likely to be overconfident, and high-ability

TABLE 3.7. Confidence, ability, and loss aversion.

	NLoss & Low			**Loss & High**	
	Yes	No		Yes	No
Overconfident	62 (53%)	56 (47%)	**Underconfident**	28 (35%)	51 (65%)

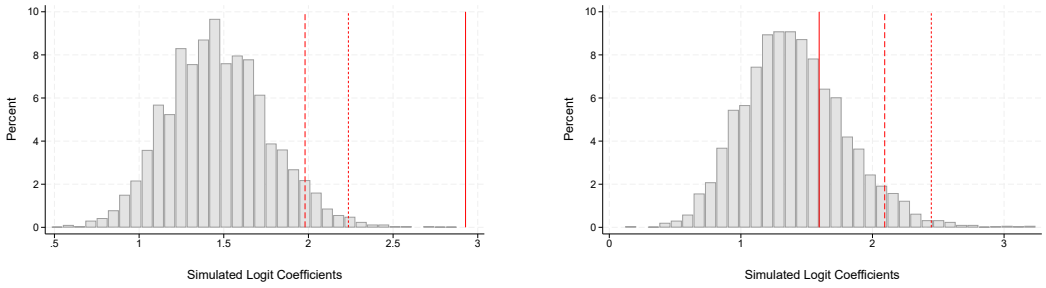
This table shows how many of overconfident (underconfident) subjects are non-loss-averse and low-ability, versus loss-averse and high-ability subjects. "Loss & High": subjects are loss-averse and high-ability (rank lower or equal to 5). "NLoss & Low": non-loss-averse with a rank above 5.

subjects – underconfident.[39] I deal with this confound by constructing a counterfactual scenario in which subjects' beliefs are drawn from the uniform distribution, whereas ranks and loss aversion parameters from their empirical distributions (for each simulation, I permute subjects' ranks and loss aversion parameters).[40] I simulate the counterfactual 5000 times, each time calculating the logit coefficient, and compare the distribution of coefficients to the one from our sample.[41] The logistic regression I estimate is of the following form:

$$\ln\left(\frac{P_{Over_i}}{1-P_{Over_i}}\right) = \alpha_0 + \alpha_1 NLA_i + \epsilon_i, \tag{3.24}$$

where P_{Over_i} is the probability that the agent is overconfident, NLA_i is an indicator variable taking value 1 if the agent is non-loss-averse and low-ability and 0 otherwise. The equation for underconfident subjects is similar, with the only difference being that the probability of interest is the probability of being underconfident, and the indicator variable takes value 1 if the agent is loss-averse and high-ability.

FIGURE 3.10. Coefficients in the simulation (gray bars) and in the data (solid red line).



(A) Overconfident (B) Underconfident

For overconfident participants, the coefficient α_1 is equal to 2.927 and highly significant (p-value = 0.000). On Panel (a) in Figure 3.10, I present a histogram of coefficients from the simulation. The dashed lines correspond to the 95^{th} and the 99^{th} percentile of the distribution. The solid red line denotes the value of the actual coefficient. It is much higher than the

[39] However, there should be no mechanical relation between loss aversion and rank or beliefs.

[40] In the baseline, I keep the matching between ranks and loss aversion as in the original dataset. The results are no different if I assign a loss aversion parameter to a rank randomly (see Appendix E.1).

[41] One consequence of drawing from empirical distributions is that the number of people who are 1) loss-averse and high-ability, and 2) non-loss-averse and low-ability, will not be the same, as neither of the two distributions is perfectly symmetric. Therefore, the distributions of logit coefficients presented in Figure 3.10 will not be the same.

simulated coefficients, meaning that the estimate could not arise as a result of the mechanical relation between overconfidence and ability. Unfortunately, we cannot say the same for underconfident participants, depicted on Panel (b) on Figure 3.10. The estimated coefficient equals 1.595 and, although it is higher than the median value in the simulation, is not high enough to pass the 95% threshold. The odds of being underconfident when one belongs to the group of high-ability, loss-averse individuals are no higher than those produced by a model with randomly assigned beliefs. Therefore, we confirm Hypothesis B.1 for overconfident but not underconfident agents.

Result B.1

Overconfident participants are more likely to be low-ability and non-loss-averse. Underconfident subjects, although more likely to be high-ability and loss-averse in absolute terms, are no more likely than they would be if their beliefs were assigned randomly.

3.4.4.2. *Prediction B.2.* In order to test the second prediction of the model, I use the following specification:

$$Bias_i = \beta_0 + \beta_1\, Over_i + \beta_2\, \lambda_i + \beta_3\, Over_i \times \lambda_i + \epsilon_i, \tag{3.25}$$

where $Bias_i$ denotes the absolute bias, $Over_i$ is an indicator variable taking value 1 if the agent is overconfident and 0 otherwise, and λ_i is the measure of loss aversion. The results are gathered in Table 3.8. As in the previous sections, I present two sets of results: estimates based on the sample of all participants (the first two columns in Table 3.8) and estimates based on the restricted sample (the last two columns in Table 3.8). The criteria for the exclusion are the same as previously. First of all, I regress the dependent variable, the absolute bias, on the indicator variable "Overconfident". The coefficient is positive and highly significant: the average absolute bias of overconfident subjects is 1.155 rank higher than the absolute bias of underconfident participants (80% increase in relative terms). This result confirms the first part of Prediction B.2.[42]

[42] As an additional exercise, I simulated Specification (1) from Table 3.8 using permutated ranks and beliefs drawn from the uniform distribution. The coefficient at the 99^{th} percentile is equal to 0.381 – a value much lower than 1.155 (the actual coefficient). This result confirms that the difference in biases between overconfident and underconfident subjects would not arise if beliefs were not subject to motivated reasoning.

TABLE 3.8. The effect of overconfidence and loss aversion on absolute bias.

Dependent variable: the absolute bias.				
	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Overconfident	1.154***	1.539**	1.155***	2.433***
	(0.230)	(0.680)	(0.230)	(0.662)
Loss Aversion		-0.041		0.021
		(0.103)		(0.099)
Overconfident × Loss Aversion		-0.080		-0.277**
		(0.132)		(0.129)
Const.	1.637***	1.837***	1.422***	1.316**
	(0.178)	(0.533)	(0.178)	(0.528)
N	227	227	197	197

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Overconfident" is an indicator variable taking value 1 if the mean of individual belief distribution was lower (better) than the actual rank.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

In the second and the last column, I gather the results of estimating (3.25). One can notice that restricting the sample has a big effect on the estimates. Again, the decisions of subjects at the boundary differ considerably from the decisions of the remaining participants. The theory provides a valid reason to exclude those participants from the analysis, hence we only interpret coefficients based on the restricted sample.

For overconfident subjects, there is a negative and significant effect of loss aversion on the absolute bias. In line with the model predictions, more loss-averse subjects end up with a smaller bias. For underconfident participants, there is no analogous effect. Although the model predicts that loss aversion should be positively correlated with the bias in this case, it also provides the reason why the effect should be much weaker than the effect for overconfident agents. This is indeed the case in our dataset: the loss aversion parameter has a stronger impact on the absolute bias for overconfident subjects compared to its effect on the bias of underconfident participants.

Result B.2

i) The absolute bias is larger for overconfident than for underconfident agents.
ii) The absolute bias is decreasing in the loss aversion parameter λ for overconfident subjects, but it is not increasing for underconfident individuals.
iii) λ has a stronger impact on the absolute bias of overconfident participants.

Results B.1 and B.2 provide additional support for the model. In both cases, the results are stronger for overconfident than for underconfident agents. It is important to note that Prediction B.2 iii) provides us with an explanation for why this could be the case. Underconfident agents tend to be high-ability, so the probability of receiving a "bad" signal is lower for them than for overconfident individuals. As a result, the loss component has a lower weight in the utility function compared to its weight in the utility function of overconfident agents, and the effect of the loss aversion parameter λ is diminished. Although the probability of receiving a "bad" signal is never zero (there is always the noise component), the effect might be hard to detect in a small sample.

3.5. Conclusions

In this paper, I present an experimental test of a model of belief choice when an agent derives utility from his beliefs. Importantly, the utility is reference-dependent, and the current belief level serves as a reference point. The model makes several predictions, most of which find confirmation in the data. The results demonstrate that, although the process of belief formation is rather complex, its outcome is far from random.

The systematic component documented in this paper is *aversion to losses* in utility derived from beliefs. In the experiment, we look at loss aversion with respect to beliefs about one's cognitive ability. The experimental data provides evidence that gain-loss attitudes in this domain drive the choice of individual beliefs. It is important to note that cognitive ability is a model example of a source of belief-based utility. The results can be generalized to any setting where an agent strongly prefers some states of the world and derives utility from believing that these states will realize. That being the case, the gain-loss attitudes should be taken into account when modeling situations such as learning about personal characteristics (or that of one's in-group/out-group), the choice of moral/selfish action in the presence of externalities, or the formation of political beliefs. As for the latter, special emphasis should

be placed on the formation of beliefs on politically contentious issues (e.g., climate change or vaccinations), since the divides on these issues are deeply rooted in personal identity, which is likely to be an important source of belief-based utility.

Moreover, as I show in the paper, loss aversion is a significant factor driving agents' bias. In this case, the relevant bias is *overconfidence* – a tendency to overestimate one's ability or performance. The data shows that participants who believe that their ability is higher than it actually is, tend to be low-ability and non-loss-averse – a result that brings us one step closer to understanding the sources of overconfidence. On a general level, this result emphasizes the importance of studying fundamental processes that govern belief formation. Unraveling the forces behind these processes could help us understand not only overconfidence but also related behavioral phenomena: information avoidance, the demand for inaccurate or slanted information, and belief polarization.

APPENDIX A

Additional Results

A.1. Results based on an indicator variable

In this section, I test Hypotheses 1-4 using a binary measure of loss aversion. I define a subject as "loss-averse" if her average response to the two hypothetical questions was above 5, which is not only the neutral value on the Likert scale we used but also the median average response. Based on this definition, there were 90 loss-averse subjects in our sample and they constituted 40% of all participants (in the restricted sample, this number dropped to 73 participants, 37% of the restricted sample). In Figure A.1, I present the average belief in the Treatment and the Control condition depending on the agent's type ("1" denotes loss-averse subjects). For non-loss-averse subjects, the average belief in Treatment is slightly above the average in Control. For loss-averse subjects, the average in Treatment is lower than in Control, in line with the model prediction.

FIGURE A.1. Beliefs of loss-averse and non-loss-averse subjects.

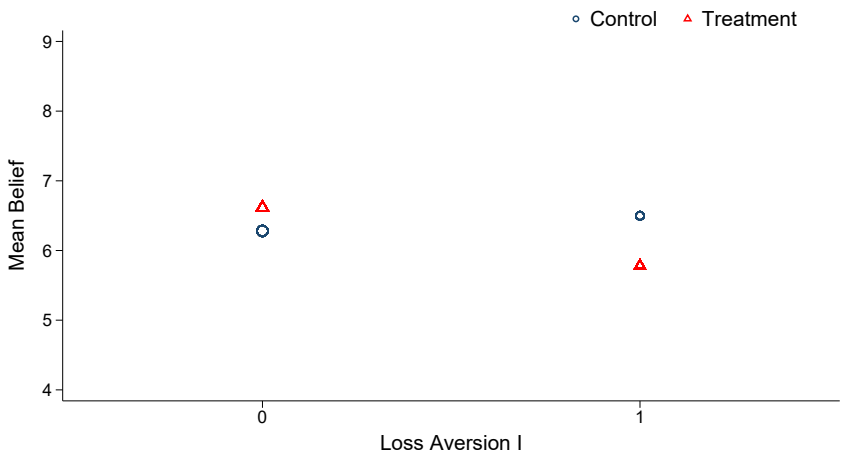


TABLE A.1. The effect of treatment and loss aversion on mean beliefs.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Treatment	0.416	0.480*	0.343	0.363
	(0.286)	(0.276)	(0.288)	(0.282)
Loss Aversion	0.355	0.355	0.761**	0.518
	(0.322)	(0.310)	(0.337)	(0.338)
Treatment × Loss Aversion	-1.326***	-1.224***	-1.589***	-1.450***
	(0.453)	(0.437)	(0.473)	(0.464)
Ability		0.169***		0.158***
		(0.040)		(0.049)
N	227	227	197	197

"Loss Aversion" is a dummy variable that takes value 1 if the average response to the two hypothetical questions was above the neutral value of 5, and 0 otherwise.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

However, the graph does not take into account the differences in ability between the two groups. For this reason, we turn to the regression analysis. In Table A.1, I present the estimates of the same regressions as in Table 3.3, with the only difference being that I replace the discrete variable "Loss Aversion" with the binary variable defined above. The coefficient at the interaction term is negative and highly significant meaning that loss-averse subjects tend to adopt lower beliefs in the Treatment condition. The results based on the indicator variable confirm Hypothesis 2.

To test Hypotheses 3 and 4, I run the regressions specified in (3.22) and (3.23) using the indicator variable. The results are gathered in Table A.2 – they correspond to the results presented in Table 3.4. The coefficients at the interaction terms in the first two columns are positive and close to significant: p-values of one-sided t-test are equal to 0.102 and 0.103 (for all participants and the restricted sample, respectively). Although the results miss the 10% significance level, they speak in favor of Hypothesis 3. Unfortunately, we cannot say the same about Hypothesis 4. The coefficient at the triple interaction is positive in the restricted

TABLE A.2. The effect of treatment, ability, and loss aversion.

	All subjects			Restricted sample (R)		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	0.016	0.072	-0.438	-0.128	-0.519	-0.320
	(0.218)	(0.513)	(0.627)	(0.230)	(0.566)	(0.658)
Loss Aversion	-0.959**	-0.205	-0.790	-1.245*	-0.089	0.280
	(0.480)	(0.552)	(0.688)	(0.662)	(0.737)	(0.961)
Ability	0.121**	0.097	0.059	0.116*	0.041	0.059
	(0.052)	(0.061)	(0.067)	(0.060)	(0.074)	(0.079)
Ability × LA	0.132	0.099	0.202*	0.176	0.119	0.059
	(0.081)	(0.081)	(0.109)	(0.107)	(0.106)	(0.146)
Treatment × LA		-1.126**	0.076		-1.551***	-2.284*
		(0.442)	(0.958)		(0.480)	(1.315)
Treatment × Ability		0.072	0.166		0.175*	0.135
		(0.080)	(0.104)		(0.098)	(0.119)
Treatment × Ability × LA			-0.228			0.127
			(0.162)			(0.213)
N	227	227	227	197	197	197

"Loss Aversion" is a dummy variable based on our measure of loss aversion. It is equal to 1 if the average response to the hypothetical questions was above 5. "Ability" is subject's position in the test score distribution (taking values from 1 to 10, with 10 assigned to subjects with the highest scores). All specifications include a constant (omitted for clarity). "All" refers to all subjects and "R" to the restricted sample. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

sample, as predicted, but far from significant (p-value of one-sided t-test = 0.550). Therefore, we cannot confirm Hypothesis 4.

The results presented in the first and the fourth column in Table A.2 shed light on Prediction 3.1. Since I do not have the measure of the cost parameter γ nor the utility u_H and u_L, I cannot directly test the prediction for non-loss-averse subjects, however, I can test the first part of Prediction 3.1, that is, the effect of p_0 (proxied by agent's ability) on beliefs of loss-averse agents. The sum of the coefficients at "Ability" and "Ability × Loss Aversion" variables is significantly larger than zero: with the p-value of a one-sided t-test = 0.000, I reject the hypothesis that the sum is lower or equal to zero. This is also true for the restricted sample. As for the second part of Prediction 3.1, all I can say is that the effect of ability for non-loss-averse subjects

is positive but lower than for loss-averse participants – the coefficient at the “Ability × Loss Aversion" variable is positive and close to significant. We conclude:

Result 3.1 (Prediction 3.1)
For loss-averse agents, an increase in ability has a positive effect on prior beliefs. For non-loss-averse agents, the effect is positive and significantly lower than for loss-averse subjects. Both results are in line with the model predictions.

In order to confirm Prediction 3.2, one has to show that the coefficient at the triple interaction is larger than the coefficient at the “Ability × Loss Aversion" variable (the case of loss-averse subjects), and the coefficient at the “Treatment × Ability” variable is lower than the coefficient at the “Ability” variable (the case of non-loss-averse agents). It is evident in Table A.2 that neither prediction holds (I confirm this observation with appropriate tests). Therefore, I cannot confirm Prediction 3.2 in our dataset.

Result 3.2 (Prediction 3.2)
For loss-averse agents, being assigned to the Treatment condition does not enhance the effect of unmanipulated beliefs (proxied by ability) on prior beliefs. For non-loss-averse agents, being assigned to the Treatment condition does not reduce this effect.

As described in the last paragraph of Section 3.4.2, this is most likely due to the treatment manipulation being not strong enough to generate an effect that would be detectable when diving sample into subgroups.

APPENDIX B

Robustness Checks

B.1. Results based on the entire sample

Before the main task, we asked participants to answer 4 control questions intended to check their understanding. In Figure B.1, I present the distribution of the number of *incorrectly* answered questions. Almost 80 % of participants answered correctly 3 or more questions, and more than 15% answered correctly 2 out of 4 questions.

In the analysis presented in the main text, we excluded 7 participants who gave incorrect answers to 3 or 4 questions. These participants constituted 3% of the initial sample. In this section, I present the results of the data analysis including those subjects. Tables B.1, B.2, and B.3 correspond to Tables 3.2, 3.3, and 3.4. One can see that the results are very similar, albeit noisier (which is to be expected, as the added subjects had problems understanding the tasks).

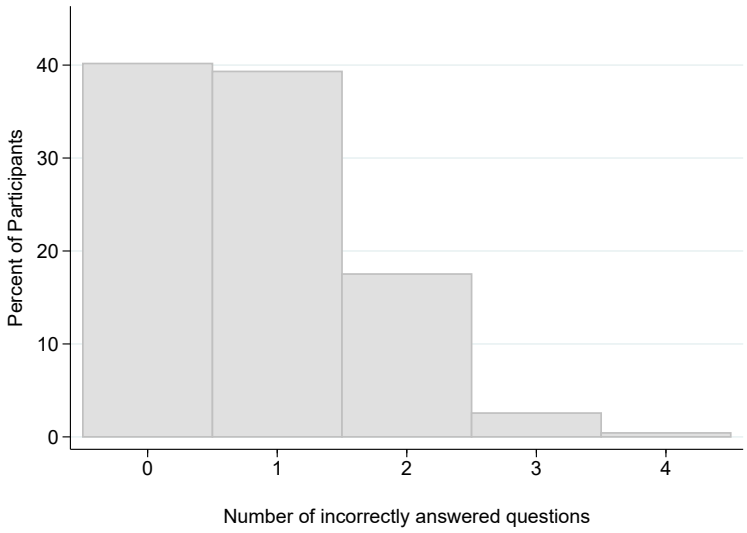


FIGURE B.1. Control Questions

TABLE B.1. The effect of treatment on mean beliefs.

	All subjects	BC	R (All - BC)
Treatment	-0.131	0.633	-0.248
	(0.225)	(0.724)	(0.235)
Const.	6.360***	5.793***	6.446***
	(0.156)	(0.503)	(0.163)
N	234	31	203

The dependent variable is the mean belief revealed in Belief Elicitation I. "All subjects" refers to the whole sample. "BC" are the boundary cases – subjects who do not fulfill the conditions necessary for the negative effect to occur (see Section 3.1.5). "R": the remaining subjects (all subjects minus boundary cases). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE B.2. The effect of treatment and loss aversion on mean beliefs.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Treatment	0.935	0.915	1.075	0.956
	(0.658)	(0.633)	(0.676)	(0.658)
Loss Aversion	-0.049	-0.050	0.037	-0.019
	(0.085)	(0.082)	(0.090)	(0.089)
Treatment × Loss Aversion	-0.216*	-0.190	-0.275**	-0.236*
	(0.128)	(0.123)	(0.133)	(0.130)
Ability		0.175***		0.174***
		(0.040)		(0.049)
N	234	234	203	203

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Ability" takes values from 1 to 10 and denotes the position in the IQ distribution. "Restricted sample" denotes sample restricted in line with theory.

TABLE B.3. The effect of treatment, ability, and loss aversion.

	All subjects			Restricted sample (R)		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	0.014	0.347	-0.582	-0.111	0.295	0.320
	(0.214)	(0.771)	(1.438)	(0.227)	(0.780)	(1.652)
Loss Aversion	-0.405***	-0.317**	-0.399**	-0.507***	-0.379**	-0.376
	(0.137)	(0.150)	(0.184)	(0.175)	(0.187)	(0.241)
Ability	-0.065	-0.101	-0.172	-0.129	-0.195	-0.193
	(0.118)	(0.123)	(0.154)	(0.142)	(0.150)	(0.194)
Ability × LA	0.050**	0.048**	0.063**	0.067**	0.066**	0.066*
	(0.023)	(0.023)	(0.030)	(0.029)	(0.029)	(0.039)
Treatment × LA		-0.168	0.022		-0.244*	-0.250
		(0.122)	(0.277)		(0.130)	(0.349)
Treatment × Ability		0.091	0.263		0.142	0.138
		(0.079)	(0.239)		(0.098)	(0.283)
Treatment × Ability × LA			-0.036			0.001
			(0.046)			(0.058)
N	234	234	234	203	203	203

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Ability" is subject's position in the test score distribution (taking values from 1 to 10, with 10 assigned to subjects with the highest test scores). All specifications include a constant (omitted for clarity). "All" refers to all participants and "R" to the restricted sample.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B.2. Results based on the median, the first and the third quartile

In this section, I present the results using an alternative dependent variable. In Tables B.4, B.5, and B.6, the analysis is based on the median of individual belief distribution (instead of the mean belief). The results are not very different from the ones presented in the main body of the paper. In Tables B.7, B.8, and B.9, I use the first and the third quartile of individual belief distribution as a dependent variable. Again, the estimated parameters are in line with the parameters from the original specification.

TABLE B.4. The effect of treatment on median beliefs.

	All subjects	BC	R (All - BC)
Treatment	-0.120	0.509	-0.219
	(0.230)	(0.750)	(0.240)
Const.	6.406***	5.813***	6.500***
	(0.160)	(0.513)	(0.167)
N	227	30	197

The dependent variable is the median belief revealed in Belief Elicitation I. "All subjects" refers to the whole sample. "BC" are the boundary cases – subjects who do not fulfill the conditions necessary for the negative effect to occur (see Section 3.1.5). "R": the remaining subjects (all subjects minus boundary cases). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE B.5. The effect of treatment and loss aversion on median beliefs.

Dependent variable: the median belief revealed in Belief Elicitation I.				
	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Treatment	0.960	0.971	1.058	0.926
	(0.671)	(0.646)	(0.684)	(0.667)
Loss Aversion	-0.051	-0.052	0.031	-0.027
	(0.087)	(0.084)	(0.092)	(0.091)
Treatment × Loss Aversion	-0.217*	-0.196	-0.266**	-0.224*
	(0.130)	(0.125)	(0.135)	(0.132)
Ability		0.175***		0.172***
		(0.041)		(0.051)
N	227	227	197	197

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Ability" takes values from 1 to 10 and denotes the position in the IQ distribution. "Restricted sample" denotes sample restricted in line with theory.

TABLE B.6. The effect of treatment, ability, and loss aversion.

	All subjects			Restricted sample (R)		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	0.039	0.315	-0.464	-0.072	0.268	0.434
	(0.219)	(0.782)	(1.466)	(0.232)	(0.794)	(1.667)
Loss Aversion	-0.445***	-0.351**	-0.416**	-0.528***	-0.403**	-0.385
	(0.139)	(0.153)	(0.185)	(0.177)	(0.190)	(0.245)
Ability	-0.101	-0.135	-0.192	-0.143	-0.207	-0.193
	(0.121)	(0.125)	(0.155)	(0.143)	(0.151)	(0.196)
Ability × LA	0.057**	0.054**	0.066**	0.070**	0.069**	0.066*
	(0.023)	(0.023)	(0.030)	(0.029)	(0.029)	(0.039)
Treatment × LA		-0.169	-0.011		-0.230*	-0.268
		(0.124)	(0.281)		(0.132)	(0.353)
Treatment × Ability		0.102	0.249		0.140	0.109
		(0.081)	(0.246)		(0.100)	(0.285)
Treatment × Ability × LA			-0.030			0.007
			(0.048)			(0.058)
N	227	227	227	197	197	197

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Ability" is subject's position in the test score distribution (taking values from 1 to 10, with 10 assigned to participants with the highest test scores). All specifications include a constant (omitted for clarity). "All" refers to the sample of all participants, and with "R" I denote the restricted sample. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE B.7. The effect of treatment on the 1^{st} and the 3^{rd} quartile.

Dependent variable:	the 1^{st} quartile	the 3^{rd} quartile
Treatment	-0.232	-0.323
	(0.232)	(0.248)
Const.	7.252^{***}	5.792^{***}
	(0.162)	(0.173)
N	197	197

The dependent variable is the 1^{st} or the 3^{rd} quartile of belief distribution revealed in Belief Elicitation I. For the sake of space, I look only at the restricted sample (all subjects minus boundary cases), for whom the model predictions should hold. Standard errors in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

TABLE B.8. The effect of loss aversion on the 1^{st} and the 3^{rd} quartile.

Dependent variable:	the 1^{st} quartile		the 3^{rd} quartile	
	(1)	(2)	(1)	(2)
Treatment	1.281^{*}	1.149^{*}	1.260^{*}	1.134
	(0.662)	(0.644)	(0.707)	(0.693)
Loss Aversion	0.086	0.028	0.110	0.054
	(0.089)	(0.088)	(0.095)	(0.095)
Treatment × Loss Aversion	-0.316^{**}	-0.274^{**}	-0.331^{**}	-0.291^{**}
	(0.130)	(0.127)	(0.139)	(0.137)
Ability		0.173^{***}		0.165^{***}
		(0.049)		(0.053)
N	197	197	197	197

The dependent variable is the 1^{st} or the 3^{rd} quartile of individual belief distribution. "Loss Aversion" takes values from 1 to 9 and denotes the average response to the two hypothetical questions. "Ability" takes values from 1 to 10 and denotes the position in the IQ distribution. Standard errors in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

TABLE B.9. The effect of treatment, ability, and loss aversion.

Dependent variable:	the 1^{st} quartile			the 3^{rd} quartile		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.083	0.520	0.741	-0.193	0.525	0.607
	(0.224)	(0.765)	(1.606)	(0.243)	(0.830)	(1.742)
Loss Aversion	-0.523***	-0.371**	-0.348	-0.424**	-0.260	-0.252
	(0.171)	(0.183)	(0.236)	(0.185)	(0.199)	(0.256)
Ability	-0.161	-0.220	-0.201	-0.103	-0.159	-0.152
	(0.138)	(0.145)	(0.188)	(0.150)	(0.158)	(0.204)
Ability × Loss Aversion	0.075***	0.072**	0.068*	0.060**	0.058*	0.056
	(0.028)	(0.028)	(0.038)	(0.030)	(0.030)	(0.041)
Treatment × Loss Aversion		-0.278**	-0.328		-0.299**	-0.317
		(0.127)	(0.340)		(0.138)	(0.369)
Treatment × Ability		0.133	0.092		0.130	0.115
		(0.097)	(0.275)		(0.105)	(0.298)
Treatment × Ability × Loss Aversion			0.009			0.003
			(0.056)			(0.061)
N	197	197	197	197	197	197

The dependent variable is the 1^{st} or the 3^{rd} quartile of individual belief distribution. "Loss Aversion" takes values from 1 to 9 and denotes the average response to the hypothetical questions. "Ability" is subject's position in the test score distribution (taking values from 1 to 10, with 10 assigned to subjects with the highest scores). All specifications include a constant (omitted for clarity). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B.3. Results based on a different restricted sample

One might wonder whether the results might be driven by a small number of people with extremely inaccurate beliefs about themselves. In Tables B.10, B.11, B.12, I show how the results from Tables 3.2, 3.3, and 3.4 change if I exclude participants with the most extreme bias. I exclude subjects 1) who were among 5% subjects with the highest bias or 5% subjects with the lowest bias (thus, I exclude participants with the most extreme positive or negative bias), and 2) whose bias exceeded the average bias ± 2 standard deviations. Because the treatment manipulation affects the variable of interest, I calculated the percentiles, the average, and the standard deviation separately for the treatment and the control group. The estimated coefficients are very similar to the ones obtained without the restrictions.

TABLE B.10. The effect of treatment on mean beliefs.

	Restricted sample I	Restricted sample II
Treatment	-0.124	-0.253
	(0.224)	(0.225)
Const.	6.457***	6.451***
	(0.155)	(0.156)
N	207	220

"Restricted sample I" denotes a sample without 5% of subjects with the highest and 5% with the lowest bias (I exclude subjects with extreme positive or negative bias). "Restricted sample II" includes only subjects whose bias did not exceed the average bias ± 2 std. The exclusion criteria were calculated separately for the treatment and control condition. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE B.11. The effect of treatment and loss aversion on mean beliefs.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	Restricted sample I		Restricted sample II	
	(1)	(2)	(1)	(2)
Treatment	1.381**	1.359**	0.894	0.850
	(0.649)	(0.578)	(0.658)	(0.614)
Loss Aversion	-0.004	0.001	-0.005	0.002
	(0.083)	(0.074)	(0.084)	(0.079)
Treatment × Loss Aversion	-0.299**	-0.264**	-0.231*	-0.200*
	(0.124)	(0.111)	(0.127)	(0.119)
Ability		0.280***		0.227***
		(0.038)		(0.039)
N	207	207	220	220

Restricted sample I and II defined as in Table B.10.

TABLE B.12. The effect of treatment, ability, and loss aversion.

	Restricted sample I			Restricted sample II		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	0.082	0.940	-1.031	-0.111	0.282	-1.350
	(0.198)	(0.727)	(1.391)	(0.208)	(0.766)	(1.454)
Loss Aversion	-0.356***	-0.222	-0.396**	-0.331**	-0.233	-0.363**
	(0.131)	(0.143)	(0.177)	(0.135)	(0.147)	(0.176)
Ability	0.065	0.057	-0.093	0.007	-0.021	-0.135
	(0.114)	(0.118)	(0.148)	(0.117)	(0.121)	(0.149)
Ability × LA	0.044**	0.039*	0.070**	0.045**	0.042*	0.066**
	(0.022)	(0.022)	(0.028)	(0.022)	(0.022)	(0.029)
Treatment × LA		-0.241**	0.151		-0.179	0.146
		(0.111)	(0.260)		(0.118)	(0.273)
Treatment × Ability		0.059	0.418*		0.089	0.389
		(0.076)	(0.229)		(0.079)	(0.241)
Treatment × Ability × LA			-0.072*			-0.061
			(0.043)			(0.046)
N	207	207	207	220	220	220

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the hypothetical questions. "Ability" is subject's position in the test score distribution. All specifications include a constant (omitted for clarity). "RS I" denotes the sample without 5% of subjects with the highest and 5% with the lowest bias (hence, I exclude subjects with extreme positive or negative bias). "RS II" includes only subjects whose bias did not exceed the average bias ± 2 std. The exclusion criteria were calculated separately for the two conditions. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B.4. Results based on a different classification of "loss-averse" subjects

In this section, I present the results of the data analysis using different ways of identifying loss-averse and non-loss-averse participants. Recall that our measure of loss aversion presented in Section 3.4.1 is ordinal, so the question of classifying subjects as loss-averse is not straightforward. Yet, to exclude the boundary cases described in Section 3.1.5, one needs to specify agents' types. In the main text, we used the average responses to the two questions as a basis for classification. Here, I present two alternative methods.

First, I classify a subject as loss-averse if his response to each of the hypothetical questions was above or equal to the median response in the sample. Using this definition 75 participants were classified as loss-averse (in the baseline, 90 were classified as loss-averse). Second, I use the average answer to the two hypothetical questions as in the baseline, but I set a lower threshold. I define a participant as loss-averse if his average answer was above 4,

a condition that gives me 136 loss-averse participants (60% of the sample) – a fraction similar to the results in Goette et al. (2019).

Importantly, we use this classification only to exclude subjects for whom the model predictions do not hold. We excluded loss-averse subjects who had the lowest ability, as well as non-loss-averse participants with the highest ability. In the baseline, we excluded 30 participants. Using the first alternative specification, this number drops to 26, but also we exclude different observations. 12 participants have a different value of the variable indicating exclusion compared to the baseline. When we use the second alternative specification, 30 participants are removed from the original sample; 10 of those have a different indicator value than in the baseline. In Tables B.13, B.14, and B.15, we present the same regressions as in Tables 3.2, 3.3, and 3.4, but exclude participants based on the new definitions of loss aversion. The results are very similar to the results presented in the main text. Therefore, we show that our results are not driven by the definition of loss aversion, and they are robust to various changes in the classification of agents' types.

TABLE B.13. The effect of treatment on mean beliefs.

	Loss Aversion I	Loss Aversion II
Treatment	-0.247	-0.121
	(0.237)	(0.235)
Const.	6.480***	6.488***
	(0.166)	(0.164)
N	201	197

Results based on data without the boundary cases, which were defined using alternative classifications. "Loss Aversion I" defines a subject as loss-averse if his responses to each of the hypothetical question were above the median response. "Loss Aversion II" assigns a status of non-loss-averse to 40% of subjects with the lowest loss aversion parameters defined as in Section 3.4.1. Standard errors in parentheses.

TABLE B.14. The effect of treatment and loss aversion on mean beliefs.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	LA I		LA II	
	(1)	(2)	(1)	(2)
Treatment	1.334*	1.200*	1.358**	1.240*
	(0.677)	(0.658)	(0.670)	(0.659)
Loss Aversion	0.064	0.014	0.071	0.021
	(0.091)	(0.090)	(0.090)	(0.090)
Treatment × Loss Aversion	-0.328**	-0.282**	-0.309**	-0.272**
	(0.133)	(0.130)	(0.132)	(0.130)
Ability		0.175***		0.144***
		(0.048)		(0.050)
N	201	201	197	197

Results based on data without the boundary cases, which were defined using alternative classifications. In "LA I", we define a subject as loss-averse if his responses to each of the two hypothetical question were above the median response. In "LA II", we assign a status of non-loss-averse to 40% of subjects with the lowest loss aversion defined in Section 3.4.1. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE B.15. The effect of treatment, ability, and loss aversion.

	LA I			LA II		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.080	0.426	0.547	0.008	0.461	0.543
	(0.228)	(0.782)	(1.615)	(0.230)	(0.773)	(1.643)
Loss Aversion	-0.512***	-0.348*	-0.335	-0.493***	-0.325*	-0.311
	(0.168)	(0.181)	(0.236)	(0.176)	(0.185)	(0.242)
Ability	-0.148	-0.215	-0.205	-0.158	-0.188	-0.227
	(0.140)	(0.146)	(0.193)	(0.141)	(0.147)	(0.192)
Ability × LA	0.071**	0.067**	0.065*	0.068**	0.063**	0.062
	(0.028)	(0.028)	(0.038)	(0.029)	(0.028)	(0.039)
Treatment × LA		-0.283**	-0.309		-0.279**	-0.339
		(0.129)	(0.332)		(0.128)	(0.348)
Treatment × Ability		0.158*	0.136		0.139	0.147
		(0.095)	(0.278)		(0.098)	(0.280)
Treatment × Ability × LA			0.005			0.009
			(0.056)			(0.057)
N	201	201	201	197	197	197

Results based on data without the boundary cases, which were defined using alternative classifications. In "LA I", we define a subject as loss-averse if his responses to each of the hypothetical questions were above the median response. In "LA II", we assign a status of non-loss-averse to 40% of subjects with the lowest loss aversion defined in Section 5.1. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

APPENDIX C

Simplified Measure of Loss Aversion II

In this section, I describe the results based on a binary measure of loss aversion that is derived from subjects' willingness to pay for signals. Since the criteria I use to classify subjects as "loss-averse" are discretionary, I present several ways in which one could define them. The results are gathered in Sections C.1 - C.4 in tables that correspond to Tables 3.2, 3.3, and 3.4 from the main text. In every table, I present two sets of results. In the first set, I exclude "boundary cases" (BC) – participants for whom the effect should not occur according to the theory.[1] In the second set, I exclude participants whose WTP for "bad" signals was positive in most decisions, indicating that they would *pay* to see "bad" signals. These decisions cannot be explained by our model. I label them as "wrong" (W) and exclude those participants from the analysis (depending on the measure of loss aversion, they constitute around 10% of the original sample).

In Section C.1, I show the results summarized in the main text (Table 3.5). The results would be almost the same if, instead of the method described in the body of the paper, I looked at the average willingness to pay for the three signals and classified a person as "loss-averse" if the average indicates paying a positive amount to avoid "bad" signals. Only 1 person would be assigned a different status compared to the baseline. For the sake of brevity, I omit these results.

In Section C.2, I conduct the same analysis as in the baseline for the five worst signals ("6", "7", "8", "9", and "10"). I classify a participant as "loss-averse" if she was willing to forgo any amount of money to not receive a "bad" signal or, in the case of monetary indifference, decided to *not* see a signal. At least 3 out of 5 signals must be avoided to be classified as "loss-averse". I classify a person as "wrong" if her WTP for "bad" signals was positive for at least 2 signals (allowing for 1 mistake). Using this definition, 34% of subjects were classified as "loss-averse" – this fraction drops to 30% if I exclude "'wrong" individuals. In Section C.3, I

[1] Whether or not a subject is in BC depends on the measure of loss aversion under consideration, therefore the number of observations in BC differs from section to section.

use the average willingness to pay for the five worst signals instead of the “at least 3 out of 5” condition. I classified a person as “loss-averse” if the average was negative. I classify a person as “wrong” if her average WTP was indicative of paying to see “bad” signals. Using this definition, 35% of subjects were classified as “loss-averse” (30% if I exclude “wrong” individuals).

Results presented in Section C.4 are based on a different definition of a “good” and a “bad” signal. I use the prior belief distribution revealed through the hypothetical choices (see Appendix D for details) and define a signal as “bad” if it was worse than the mean of this distribution. I classify a participant as “loss-averse" if the average willingness to pay for (individually defined) “bad” signals indicated information avoidance. I classify a person as “wrong” if her average WTP for “bad” signals indicated that she wanted to see these signals, contrary to the model predictions.

The results are very similar across the four specifications. I compare them to the results described in Appendix A, which are also based on a binary variable (in the main text I use the discrete 9-point Likert scale). Although the main effect – the coefficient at the interaction of “Treatment” and “Loss aversion” – is 35-60% lower than the effect based on the first measure (see Table A.1) and passes the threshold for significance only in Section C.2, in all specifications it goes in the predicted direction. Moreover, in the saturated model (the last columns in the last table), the interaction “Treatment × Loss Aversion” is negative and significant, as predicted by the theory.

Interestingly, Prediction 3.2, which is confirmed with the first measure of loss aversion (see Result 3.2 in Appendix A), finds some confirmation in the results based on the second measure. Although we cannot confirm it with sufficient confidence (the p-values range from 0.191 to 0.220, depending on the specification), the coefficient at the triple interaction is larger than the coefficient at the “Ability × Loss Aversion" variable, in line with the first part of Prediction 3.2 (the case of loss-averse subjects). At the same time, the coefficient at the “Treatment × Ability” variable is lower than the coefficient at the “Ability” variable in every specification, although it also does not pass the 10%-significance level (the p-values range from 0.175 to 0.252). This result is in line with the second part of Prediction 3.2 (the case of non-loss-averse agents).

C.1. Results based on coarse classification (three worst signals)

TABLE C.1. The effect of treatment on mean beliefs.

	R1 (All - BC)	R2 (BC)	R3 (All - BC - W)
Treatment	-0.349	1.444	-0.191
	(0.229)	(0.963)	(0.245)
Const.	6.575***	5.549***	6.522***
	(0.164)	(0.472)	(0.176)
N	202	25	178

"R1" denotes the sample without the boundary cases – subjects who do not fulfill the conditions necessary for negative effect to occur (based on the new measure of loss aversion). "R2" includes only the boundary cases. In "R3", I also exclude subjects whose decisions cannot be explained by the model: those willing to pay to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.2. The effect of treatment and loss aversion.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	R1 (All - BC)		R3 (All - BC - W)	
	(1)	(2)	(1)	(2)
Treatment	-0.207	-0.033	0.046	0.198
	(0.267)	(0.262)	(0.292)	(0.288)
Loss Aversion	0.379	0.222	0.475	0.340
	(0.373)	(0.363)	(0.384)	(0.375)
Treatment × Loss Aversion	-0.541	-0.625	-0.793	-0.857
	(0.521)	(0.505)	(0.536)	(0.521)
Ability		0.181***		0.178***
		(0.048)		(0.052)
N	202	202	178	178

"R1" denotes a sample without the boundary cases: those who don't fulfill conditions necessary for the negative effect to occur (based on the new measure of loss aversion). In "R3", I also exclude subjects willing to pay to see a negative signal.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.3. The effect of treatment, ability, and loss aversion.

	RS1			RS3		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.211	-0.690	-0.434	-0.074	-0.148	0.364
	(0.226)	(0.559)	(0.640)	(0.242)	(0.630)	(0.745)
Loss Aversion	-0.802	-0.402	0.149	-0.922	-0.389	0.518
	(0.671)	(0.751)	(1.008)	(0.701)	(0.784)	(1.055)
Ability	0.145**	0.086	0.113	0.128**	0.106	0.158*
	(0.056)	(0.076)	(0.083)	(0.064)	(0.086)	(0.095)
Ability × LA	0.120	0.109	0.021	0.140	0.121	-0.025
	(0.105)	(0.106)	(0.152)	(0.110)	(0.111)	(0.159)
Treatment × LA		-0.688	-1.715		-0.836	-2.513*
		(0.518)	(1.355)		(0.534)	(1.415)
Treatment × Ability		0.122	0.072		0.059	-0.038
		(0.096)	(0.113)		(0.105)	(0.129)
Treatment × Ability × LA			0.174			0.284
			(0.212)			(0.222)
N	202	202	202	178	178	178

"R1" denotes the sample without the boundary cases – those who do not fulfill the conditions necessary for negative effect to occur (based on the second measure of loss aversion). In "R3", I also exclude subjects whose decisions cannot be explained by the model: to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C.2. Results based on coarse classification (five worst signals)

TABLE C.4. The effect of treatment on mean beliefs.

	R1 (All - BC)	R2 (BC)	R3 (All - BC - W)
Treatment	-0.338	1.302	-0.164
	(0.229)	(0.959)	(0.253)
Const.	6.571***	5.570***	6.531***
	(0.165)	(0.470)	(0.181)
N	202	25	172

"R1" denotes the sample without the boundary cases – subjects who do not fulfill the conditions necessary for negative effect to occur (based on the new measure of loss aversion). "R2" includes only the boundary cases. In "R3" I also exclude subjects whose decisions cannot be explained by the model: those willing to pay to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.5. The effect of treatment and loss aversion.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	R1 (All - BC)		R3 (All - BC - W)	
	(1)	(2)	(1)	(2)
Treatment	-0.206	-0.037	0.068	0.210
	(0.265)	(0.259)	(0.300)	(0.293)
Loss Aversion	0.485	0.341	0.566	0.452
	(0.389)	(0.376)	(0.406)	(0.393)
Treatment × Loss Aversion	-0.554	-0.685	-0.817	-0.942*
	(0.530)	(0.511)	(0.558)	(0.540)
Ability		0.193***		0.196***
		(0.047)		(0.054)
N	202	202	172	172

"R1" denotes a sample without the boundary cases: those who don't fulfill conditions necessary for the negative effect to occur (based on the new measure of loss aversion). In "R3", I also exclude subjects willing to pay to see a negative signal.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.6. The effect of treatment, ability, and loss aversion.

	RS1			RS3		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.232	-0.759	-0.553	-0.092	-0.225	0.284
	(0.224)	(0.550)	(0.629)	(0.248)	(0.652)	(0.775)
Loss Aversion	-0.846	-0.369	0.100	-0.984	-0.414	0.474
	(0.669)	(0.741)	(1.015)	(0.731)	(0.800)	(1.084)
Ability	0.150***	0.087	0.108	0.136**	0.107	0.156
	(0.056)	(0.074)	(0.080)	(0.067)	(0.087)	(0.096)
Ability × LA	0.140	0.125	0.048	0.159	0.146	0.000
	(0.105)	(0.105)	(0.154)	(0.114)	(0.114)	(0.165)
Treatment × LA		-0.779	-1.622		-0.964*	-2.608*
		(0.521)	(1.348)		(0.550)	(1.464)
Treatment × Ability		0.134	0.094		0.074	-0.021
		(0.094)	(0.112)		(0.108)	(0.133)
Treatment × Ability × LA			0.143			0.276
			(0.210)			(0.228)
N	202	202	202	172	172	172

"R1" denotes the sample without the boundary cases – those who do not fulfill the conditions necessary for negative effect to occur (based on the second measure of loss aversion). In "R3", I also exclude subjects whose decisions cannot be explained by the model: to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C.3. Results based on coarse classification (five worst signals, average)

TABLE C.7. The effect of treatment on mean beliefs.

	R1 (All - BC)	R2 (BC)	R3 (All - BC - W)
Treatment	-0.338	1.353	-0.230
	(0.228)	(0.983)	(0.256)
Const.	6.570***	5.518***	6.533***
	(0.163)	(0.491)	(0.182)
N	203	24	166

"R1" denotes the sample without the boundary cases – subjects who do not fulfill the conditions necessary for negative effect to occur (based on the new measure of loss aversion). "R2" includes only the boundary cases. In "R3" I also exclude subjects whose decisions cannot be explained by the model: those willing to pay to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.8. The effect of treatment and loss aversion.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	R1 (All - BC)		R3 (All - BC - W)	
	(1)	(2)	(1)	(2)
Treatment	-0.223	-0.056	-0.029	0.095
	(0.266)	(0.260)	(0.311)	(0.304)
Loss Aversion	0.384	0.206	0.470	0.314
	(0.372)	(0.361)	(0.392)	(0.384)
Treatment × Loss Aversion	-0.437	-0.558	-0.631	-0.711
	(0.521)	(0.502)	(0.551)	(0.535)
Ability		0.192***		0.182***
		(0.047)		(0.055)
N	203	203	166	166

"R1" denotes a sample without the boundary cases: those who don't fulfill conditions necessary for the negative effect to occur (based on the new measure of loss aversion). In "R3", I also exclude subjects willing to pay to see a negative signal.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.9. The effect of treatment, ability, and loss aversion.

	RS1			RS3		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.221	-0.827	-0.554	-0.153	-0.442	0.121
	(0.223)	(0.547)	(0.626)	(0.251)	(0.654)	(0.786)
Loss Aversion	-0.829	-0.456	0.137	-0.984	-0.553	0.366
	(0.674)	(0.740)	(0.992)	(0.726)	(0.797)	(1.070)
Ability	0.153***	0.083	0.111	0.121*	0.079	0.135
	(0.056)	(0.074)	(0.080)	(0.068)	(0.089)	(0.099)
Ability × LA	0.126	0.118	0.022	0.158	0.148	-0.002
	(0.104)	(0.104)	(0.149)	(0.113)	(0.113)	(0.163)
Treatment × LA		-0.687	-1.807		-0.752	-2.483*
		(0.515)	(1.351)		(0.548)	(1.456)
Treatment × Ability		0.145	0.092		0.095	-0.012
		(0.094)	(0.111)		(0.109)	(0.137)
Treatment × Ability × LA			0.187			0.291
			(0.208)			(0.226)
N	203	203	203	166	166	166

"R1" denotes the sample without the boundary cases – those who do not fulfill the conditions necessary for negative effect to occur (based on the second measure of loss aversion). In "R3", I also exclude subjects whose decisions cannot be explained by the model: to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C.4. Results based on coarse classification (the relative measure)

TABLE C.10. The effect of treatment on mean beliefs.

	R1 (All - BC)	R2 (BC)	R3 (All - BC - W)
Treatment	-0.357	1.550	-0.239
	(0.228)	(0.916)	(0.258)
Const.	6.570***	5.518***	6.519***
	(0.163)	(0.485)	(0.184)
N	202	25	164

"R1" denotes the sample without the boundary cases – subjects who do not fulfill the conditions necessary for negative effect to occur (based on the new measure of loss aversion). "R2" includes only the boundary cases. In "R3" I also exclude subjects whose decisions cannot be explained by the model: those willing to pay to see a negative signal. Standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.11. The effect of treatment and loss aversion.

Dependent variable: the mean belief revealed in Belief Elicitation I.				
	R1 (All - BC)		R3 (All - BC - W)	
	(1)	(2)	(1)	(2)
Treatment	-0.224	-0.070	-0.013	0.103
	(0.263)	(0.257)	(0.308)	(0.302)
Loss Aversion	0.389	0.202	0.492	0.330
	(0.381)	(0.371)	(0.402)	(0.394)
Treatment × Loss Aversion	-0.550	-0.603	-0.761	-0.780
	(0.534)	(0.516)	(0.564)	(0.548)
Ability		0.186***		0.178***
		(0.047)		(0.055)
N	202	202	164	164

"R1" denotes a sample without the boundary cases: those who don't fulfill conditions necessary for the negative effect to occur (based on the new measure of loss aversion). In "R3", I also exclude subjects willing to pay to see a negative signal.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.12. The effect of treatment, ability, and loss aversion.

	RS1			RS3		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.226	-0.801	-0.550	-0.139	-0.349	0.194
	(0.224)	(0.551)	(0.625)	(0.254)	(0.663)	(0.790)
Loss Aversion	-0.800	-0.407	0.154	-0.932	-0.455	0.456
	(0.686)	(0.766)	(1.012)	(0.739)	(0.824)	(1.093)
Ability	0.153***	0.086	0.112	0.126*	0.091	0.145
	(0.055)	(0.074)	(0.080)	(0.068)	(0.090)	(0.100)
Ability × LA	0.117	0.108	0.019	0.146	0.132	-0.015
	(0.106)	(0.107)	(0.150)	(0.115)	(0.116)	(0.164)
Treatment × LA		-0.694	-1.775		-0.780	-2.519*
		(0.529)	(1.381)		(0.562)	(1.487)
Treatment × Ability		0.138	0.089		0.079	-0.023
		(0.095)	(0.111)		(0.111)	(0.137)
Treatment × Ability × LA			0.181			0.293
			(0.214)			(0.232)
N	202	202	202	164	164	164

"R1" denotes the sample without the boundary cases – those who do not fulfill the conditions necessary for negative effect to occur (based on the second measure of loss aversion). In "R3", I also exclude subjects whose decisions cannot be explained by the model: to see a negative signal. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

APPENDIX D

Extended Measure of Loss Aversion II

One can derive the measure of loss aversion in two steps. First, I use the hypothetical choices to obtain a posterior belief distribution. Then, I use the posteriors and subjects' willingness to pay to avoid signals to retrieve the individual measure of loss aversion λ.

Posterior Beliefs. For every $x \in \{1, ..., 10\}$, participants report a probability p^x_{B2} that the signal x would come from Box 2. I assume that subjects report their beliefs truthfully and their beliefs about the box correspond to the posterior beliefs about the signal.[1] I argue that this assumption is not very demanding considering our signal structure. For example, if an agent reports that, after seeing "5", the probability of the ball being drawn from Box 2 is 70%, I assume that she would conclude that with probability 70% the number "5" is her rank and with probability 30% it is not. Because the person believes that with probability 30% the number came from the entirely uninformative Box 1, I assume that she places the remaining 30% probability on her prior. Let us denote the posterior probability after receiving a signal x with a 10-element vector p^x_1 (each element of the vector corresponds to the probability placed on one of the 10 ranks). I calculate the posterior using the following formula:

$$p^x_1 = p^x_{B2} \cdot e_{i=x} + (1 - p^x_{B2}) \cdot p_0 \tag{D.1}$$

where $e_{i=x}$ is a null vector with one entry equal to 1 at x, and p_0 is the vector of prior probabilities. I derive the latter using the subjects' decisions about all 10 numbers. Intuitively, if a participant concludes that, after seeing number "5", it is his rank with probability 60%, and after seeing number "4", it is his rank with probability 30%, then he places twice as high probability on the number being "5" in his prior beliefs. With the decisions on all 10 numbers, one can back out the prior beliefs distribution that they use for their choices. Importantly, I derive p_0 and p^x_1 using a different method than the stated belief $\tilde{p}_0$. Although I have no direct proof that p_0 and p^x_1 are the "unmanipulated" belief, there is some evidence that people

[1] The beliefs in Period 0 about the posterior that the agent will form in Period 1 might be different from the actual posterior she forms in Period 1. However, since the utility U_0 in equation (3.20) is evaluated by the agent in Period 0, it is not inappropriate to use her beliefs in Period 0 about the posterior she will form in Period 1.

reveal beliefs that are closer to the truth in hypothetical choices (Kozakiewicz, 2020). In what follows, I use beliefs revealed from hypothetical choices as a measure of the unmanipulated posterior belief in (3.20).

Reducing the State Space. In the experiment, subjects are learning about the state of the world that can take one of ten values (recall that the rank is an integer between 1 and 10). In order to test the mechanism of the model in our dataset, I reduce the state space as follows. I define a "high" state to be a state that is better than or equal to the median of the distribution revealed through hypothetical choices. A "good" signal is a signal that indicates a "high" state – a signal that is better than or equal to the median belief (a "low" state and a "bad" signal are defined in the same way).

The posterior p_1^H is defined as an average of the posterior probabilities that the agent form after signals that are lower (better) than the median. I use the following formula:

$$p_1^H = \frac{1}{|S|} \sum_{x \in S} \sum_{s \leq \bar{m}} p_{1,s}^x \tag{D.2}$$

where S denotes the set of all "good" signals. With $p_{1,s}^x$ I denote the posterior probability assigned to the state s after a signal x (the s-th element of p_1^x, the vector of posterior beliefs after a signal x). The inner sum calculates the total probability that a subject assigns, after a signal x, to the ranks that are lower (better) than his median belief $\bar{m}$. In other words, this is the probability assigned to the state being "high" after a signal x. The outer sum, divided by the number of elements of S, gives us the average posterior belief assigned the "high" state (the average is calculated over all "good" signals). The average posterior probability after a "bad" signal, p_1^L, is defined analogously.

Example 1. Consider an agent whose point allocation is summarized in a vector: $p_{B2} = [0,0,30,40,60,40,20,0,0,0]$ (the n-th element of the vector corresponds to the decision considering a signal n). Using his decisions about the boxes, I obtain his prior belief vector $p_0 = [0,0,0.16,0.21,0.32,0.21,0.1,0,0,0]$ (the n-th element of the vector corresponds to the probability assigned to rank n). One can note that the agent perceives signals below or equal to "5" to be "good" signals. I use the same choices to construct conditional posterior probabilities: after receiving a signal "3", the agents would form a posterior $p_1^{x=3} = [0,0,0.41,0.14,0.22,0.15,0.08,0,0,0]$; after receiving a signal "4", he would form a posterior $p_1^{x=4} = [0,0,0.1,0.52,0.19,0.12,0.07,0,0,0]$; after a signal "5", it would be $p_1^{x=5} = [0,0,0.065,0.085,0.72,0.085,0.045,0,0,0]$, and so on.

Hence, after a signal "3" the agent would assign 77% probability that the state is high (his rank is below or equal to 5), after a signal "4" this probability would be 81%, and after a signal "5" it would be equal to 87%. After a signal "1" or "2", the agent's posterior would be equal to his prior p_0, and the probability assigned to the state being high would be 67%. The average posterior probability p_1^H is therefore $\frac{1}{5}(0.67+0.67+0.77+0.81+0.87) = 0.758$.

Monetary Equivalents. Subject's choices in the price lists provide us with a monetary equivalent δ_x for every signal $x \in \{1,...,10\}$. I reduce the state space in the same way as for the posterior beliefs. The monetary equivalent of receiving a "high" signal is the average taken over the monetary equivalents for "good" signals:

$$\delta_H = \frac{1}{|S|}\sum_{x\in S}\delta_x \tag{D.3}$$

where S denotes the set of all "good" signals, that is, signals indicating a state that is better than or equal to the median belief, and δ_x is the highest amount of money the participant is willing to pay to avoid a signal x. The monetary equivalent of a "low" signal, δ_L, is defined analogously.

As for δ_x, I derive it from subjects' decisions in price lists as follows. I assume a subject's switching point to be the midpoint between the two decisions in which the participant switched from Option B to Option A. For example, if the agent chooses Option A over Option B when the monetary transfer in Option B is lower or equal to 1.70 Euro, but decides to take Option A when he gets 1.80 Euro or more, I assume that his switching point is 1.75 Euro. I assume that the agent is willing to forgo 25 cents to avoid seeing the signal and I assign him $\delta_x = -0.25$ for the signal x. In the case of participants whose switching points are not captured by the list (e.g., a person who chose Option A in every decision), I assume the monetary equivalent to be $\delta_x = -1.05$ or $\delta_x = 1.05$.

Loss Aversion Measure. Having constructed the posterior beliefs p_1^H and p_1^L, as well as the monetary equivalents δ_H and δ_L, one can derive the individual measure of loss aversion in the following way. From the equation (3.20), we have:

$$\delta_H = p_1^H u_H + (1-p_1^H)u_L + \eta\left[p_1^H u_H + (1-p_1^H)u_L - \Big(p_0 u_H + (1-p_0)u_L\Big)\right], \tag{D.4}$$

$$\delta_L = p_1^L u_H + (1-p_1^L)u_L + \eta\lambda\left[p_1^L u_H + (1-p_1^L)u_L - \Big(p_0 u_H + (1-p_0)u_L\Big)\right]. \tag{D.5}$$

After setting $u_L = 1$ and assuming $\eta = 1$, we have two equations with two unknowns, u_H and λ. By rearranging the first equation in (D.4), we get the formula for u_H:

$$u_H = \left[\delta_H + (2p_1^H - p_0)\right] \frac{1}{2p_1^H - p_0} = 1 - \frac{1-\delta_H}{2p_1^H - p_0}.$$

The second equation, if we set $u_L = 1$, $\eta = 1$ and rearrange, takes the form:

$$\delta_L = p_1^L u_H + (1 - p_1^L) + \lambda(p_1^L - p_0)(u_H - 1),$$

which can be further rearranged:

$$\lambda = \frac{\delta_L - p_1^L u_H + p_1^L - 1}{(u_H - 1)(p_1^L - p_0)}.$$

After substituting u_H derived from the first equation, we get the following formula:

$$\lambda = \frac{(\delta_L - 1)(2p_1^H - p_0)}{(\delta_H - 1)(p_1^L - p_0)} - \frac{p_1^L}{p_1^L - p_0}, \tag{D.6}$$

where p_0 is the probability of the state being "high" revealed through the decision about the boxes, whereas p_1^H and p_1^L are the average posterior probabilities after a "good" and a "bad" signal, respectively.

An important question is whether agents take into account the utility from the future belief level – the first two terms in (D.4) – when evaluating the prospect. One could imagine that, for various reasons (e.g., imperfect attention or difficulties in hypothetical thinking), the agent might only consider the shift represented by the gain-loss component.[2] If this was the case, the equations for δ_H and δ_L would take the form:

$$\delta_H = \eta\left[p_1^H u_H + (1 - p_1^H)u_L - \left(p_0 u_H + (1 - p_0)u_L\right)\right], \tag{D.7}$$

$$\delta_L = \eta\lambda\left[p_1^L u_H + (1 - p_1^L)u_L - \left(p_0 u_H + (1 - p_0)u_L\right)\right]. \tag{D.8}$$

After taking the same steps as above (solving the first equation for u_H, which is then substituted into the second equation), we get the formula:

$$\lambda = \frac{\delta_L}{\delta_H} \frac{p_1^H - p_0}{p_1^L - p_0}, \tag{D.9}$$

[2] Using this assumption when specifying the utility function (1), would not change the comparative statics nor the model predictions. Which specification is better in describing the decision problem remains an open question.

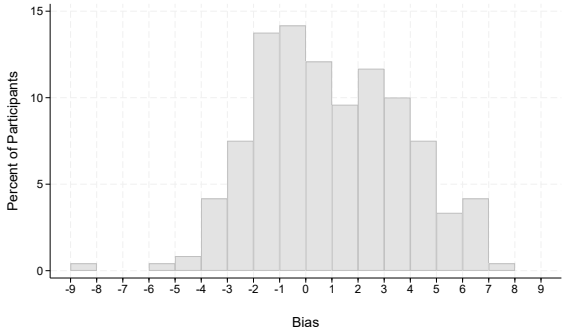
which is more straightforward than (D.6). Intuitively, the measure of loss aversion depends on the amount of money a subject is willing to forgo to avoid "bad" news (relative to the amount of money he would pay to avoid "good" news) weighted by how much his beliefs would move after each signal.

APPENDIX E

Over- and Underconfidence

In this section, I describe the results of an additional simulation, in which subjects' ranks and loss aversion parameters were no longer matched like in the data. I start by describing overconfident and underconfident participants in more detail. I classify a subject as overconfident if the mean of belief distribution he reported in Belief Elicitation I was *higher* than his actual position (using the reversed variables, which I described in Section 3.4.1). The remaining subjects are classified as underconfident. There were 135 overconfident and 92 underconfident subjects in our sample.[1] The distribution of subjects' bias is presented in Figure E.1. The average bias of overconfident participants was equal to 2.79 (the standard deviation was equal to 1.89), whereas the average bias of underconfident participants was -1.64 (the standard deviation was 1.37). As expected, the two types differ in terms of cognitive ability. The average position of underconfident agents was equal to 7.54 (expressed in deciles of the IQ distribution), and that of overconfident participants was 3.84. The difference in the average levels of loss aversion arises only in the restricted sample. The average response to the hypothetical questions was equal to 4.58 for overconfident subjects, a value that was lower than the average revealed by underconfident subjects (the difference of 0.43 is significant at the 5% level).

FIGURE E.1. Distribution of subjects' bias (based on Belief Elicitation I).

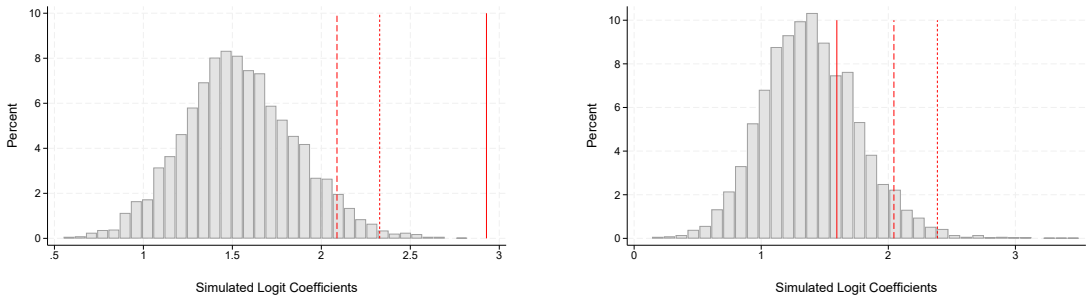


[1] One participant revealed a mean belief equal to his position. I classify him as underconfident, however, assigning him to the other group does not change the results.

E.1. Simulation (robustness check)

The results described in this section complement the results of a simulation described in Section 3.4.4.1. The simulation shown in the main text involved 1) random assignment of beliefs about rank, and 2) permutation of the distribution of actual ranks and loss aversion parameters *without breaking the matching* between the two. If a participant i's rank was 5 and he revealed a loss aversion parameter equal to 4, this participant will keep his rank and loss aversion in the simulated data set, and only his belief (thus, also the confidence type) will be assigned at random in each iteration. In Figure E.2, I show the results of a simulation without the empirical relation between the subject's rank and loss aversion parameter. For each participant, I randomly draw belief from the uniform distribution *and* the loss aversion parameter from the empirical distribution. The results are very similar to the ones presented in Figure 3.10. For overconfident agents, the actual coefficient is much higher than any estimated value, including the values indicating 95^{th} and 99^{th} percentile of the distribution (denoted with dashed red lines on the graph). For underconfident agents, the coefficient based on the actual data does not pass the 5% threshold. Therefore, I confirm that the results presented in Section 3.4.4.1 are robust to the described modification, hence they are not driven by the relationship between ability and loss aversion.

FIGURE E.2. Simulated coefficients (gray bars) and estimated (solid red line).



(A) Overconfident (B) Underconfident

Bibliography

Abadie, Alberto and Guido W Imbens (2006). "Large sample properties of matching estimators for average treatment effects". In: *Econometrica* 74.1, pp. 235–267.

Abdellaoui, Mohammed, Philippe Colo, and Brian Hill (2021). "Eliciting multiple prior beliefs". In: *HEC Paris Research Paper No. ECO/SCD-2021-1426*.

Akerlof, George A and William T Dickens (1982). "The economic consequences of cognitive dissonance". In: *American Economic Review* 72.3, pp. 307–319.

Ambuehl, Sandro and Shengwu Li (2018). "Belief updating and the demand for information". In: *Games and Economic Behavior* 109, pp. 21–39.

Barber, Brad M and Terrance Odean (2001). "Boys will be boys: Gender, overconfidence, and common stock investment". In: *The Quarterly Journal of Economics* 116.1, pp. 261–292.

Barron, Kai (2021). "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?" In: *Experimental Economics* 24.1, pp. 31–58.

Bell, David E (1985). "Disappointment in decision making under uncertainty". In: *Operations Research* 33.1, pp. 1–27.

Bénabou, Roland and Jean Tirole (2016). "Mindful Economics: The Production, Consumption, and Value of Beliefs". In: *Journal of Economic Perspectives* 30.3, pp. 141–164.

Benjamin, Daniel J (2019). "Errors in probabilistic reasoning and judgment biases". In: *Handbook of Behavioral Economics: Applications and Foundations 1* 2, pp. 69–186.

Bracha, Anat and Donald J Brown (2012). "Affective decision making: A theory of optimism bias". In: *Games and Economic Behavior* 75.1, pp. 67–80.

Brandts, Jordi and Gary Charness (2009). "The strategy method: A survey of experimental evidence". Working Paper.

Brunnermeier, Markus K and Jonathan A Parker (2005). "Optimal expectations". In: *American Economic Review* 95.4, pp. 1092–1118.

Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini (2013). "Overconfidence and social signalling". In: *The Review of Economic Studies* 80.3, pp. 949–983.

Buser, Thomas, Leonie Gerhards, and Joël Van Der Weele (2018). "Responsiveness to feedback as a personal trait". In: *Journal of Risk and Uncertainty* 56.2, pp. 165–192.

Bushong, Benjamin and Tristan Gagnon-Bartsch (2023). "Reference dependence and attribution bias: evidence from real-effort experiments". In: *American Economic Journal: Microeconomics* 15.2, pp. 271–308.

Camerer, Colin and Dan Lovallo (1999). "Overconfidence and Excess Entry: An Experimental Approach". In: *American Economic Review* 89.1, pp. 306–318.

Caplin, Andrew and John Leahy (2001). "Psychological expected utility theory and anticipatory feelings". In: *The Quarterly Journal of Economics* 116.1, pp. 55–79.

Caplin, Andrew and John V Leahy (2019). "Wishful Thinking". Working Paper.

Carroll, Patrick, Kate Sweeny, and James A Shepperd (2006). "Forsaking optimism". In: *Review of General Psychology* 10.1, pp. 56–73.

Charness, Gary and Dan Levin (2009). "The origin of the winner's curse: a laboratory study". In: *American Economic Journal: Microeconomics* 1.1, pp. 207–36.

Chew, Soo Hong, Wei Huang, and Xiaojian Zhao (2020). "Motivated false memory". In: *Journal of Political Economy* 128.10, pp. 3913–3939.

Coutts, Alexander (2019). "Good news and bad news are still news: Experimental evidence on belief updating". In: *Experimental Economics* 22.2, pp. 369–395.

Coutts, Alexander, Leonie Gerhards, and Zahra Murad (2020). "What to blame? Self-serving attribution bias with multi-dimensional uncertainty". Working Paper.

Drobner, Christoph (2022). "Motivated beliefs and anticipation of uncertainty resolution". In: *American Economic Review: Insights* 4.1, pp. 89–105.

Drobner, Christoph and Sebastian Goerg (2021). "Motivated belief updating and rationalization of information". Working Paper.

Eil, David and Justin M. Rao (2011). "The good news-bad news effect: Asymmetric processing of objective information about yourself". In: *American Economic Journal: Microeconomics* 3.2, pp. 114–138.

Engelmann, Jan, Maël Lebreton, Peter Schwardmann, Joel J van der Weele, and Li-Ang Chang (2019). "Anticipatory anxiety and wishful thinking". Working Paper.

Enke, Benjamin (2020). "What you see is all there is". In: *The Quarterly Journal of Economics* 135.3, pp. 1363–1398.

Enke, Benjamin and Florian Zimmermann (2017). "Correlation neglect in belief formation". In: *The Review of Economic Studies* 86.1, pp. 313–332.

Ertac, Seda (2011). "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback". In: *Journal of Economic Behavior & Organization* 80.3, pp. 532–545.

Esponda, Ignacio and Emanuel Vespa (2014). "Hypothetical thinking and information extraction in the laboratory". In: *American Economic Journal: Microeconomics* 6.4, pp. 180–202.

— (2016). "Contingent preferences and the sure-thing principle: Revisiting classic anomalies in the laboratory". Working Paper.

— (2018). "Endogenous sample selection: A laboratory study". In: *Quantitative Economics* 9.1, pp. 183–216.

Eyster, Erik and Matthew Rabin (2014). “Extensive imitation is irrational and harmful”. In: *The Quarterly Journal of Economics* 129.4, pp. 1861–1898.

Falk, Armin and Florian Zimmermann (2017). “Consistency as a signal of skills”. In: *Management Science* 63.7, pp. 2197–2210.

Fischbacher, Urs (2007). “z-Tree: Zurich toolbox for ready-made economic experiments”. In: *Experimental Economics* 10.2, pp. 171–178.

Fudenberg, Drew and David K Levine (2006). “A dual-self model of impulse control”. In: *American Economic Review* 96.5, pp. 1449–1476.

Gagnon-Bartsch, Tristan and Benjamin Bushong (2022). “Learning with misattribution of reference dependence.” In: *Journal of Economic Theory* 203, p. 105473.

Gerlitz, Jean-Yves and Jürgen Schupp (2005). “Zur Erhebung der Big-Five-basierten persoenlichkeitsmerkmale im SOEP”. In: *DIW Research Notes* 4.

Goette, Lorenz, Thomas Graeber, Alexandre Kellogg, and Charles Sprenger (2019). “Heterogeneity of gain-loss attitudes and expectations-based reference points”. Working Paper.

Gollier, Christian and Alexander Muermann (2010). “Optimal choice and beliefs with ex ante savoring and ex post disappointment”. In: *Management Science* 56.8, pp. 1272–1284.

Golman, Russell, David Hagmann, and George Loewenstein (2017). “Information avoidance”. In: *Journal of Economic Literature* 55.1, pp. 96–135.

Gross, James J and Oliver P John (2003). “Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being.” In: *Journal of Personality and Social Psychology* 85.2, p. 348.

Grossman, Zachary and David Owens (2012). “An unlucky feeling: Overconfidence and noisy feedback”. In: *Journal of Economic Behavior & Organization* 84.2, pp. 510–524.

Hanna, Rema, Joshua Schwartzstein, and Sendhil Mullainathan (2014). “Learning Through Noticing: Theory and Evidence From a Field Experiment”. In: *The Quarterly Journal of Economics*, pp. 1311–1353.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack (2018). “Unrealistic expectations and misguided learning”. In: *Econometrica* 86.4, pp. 1159–1214.

Hestermann, Nina and Yves Le Yaouanq (2021). “Experimentation with self-serving attribution biases”. In: *American Economic Journal: Microeconomics* 13.3, pp. 198–237.

Hossain, Tanjim and Ryo Okui (2013). “The binarized scoring rule”. In: *The Review of Economic Studies* 80.3, pp. 984–1001.

Huffman, David, Collin Raymond, and Julia Shvets (2022). “Persistent overconfidence and biased memory: Evidence from managers”. In: *American Economic Review* 112.10, pp. 3141–3175.

Kőszegi, Botond (2006). “Ego Utility, Overconfidence, and Task Choice”. In: *Journal of the European Economic Association* 4.June, pp. 673–707.

Kőszegi, Botond (Sept. 2010). "Utility from anticipation and personal equilibrium". In: *Economic Theory* 44.3, pp. 415–444.

Kőszegi, Botond and Matthew Rabin (2006). "A model of reference-dependent preferences". In: *The Quarterly Journal of Economics* 121.4, pp. 1133–1165.

— (2009). "Reference-dependent consumption plans". In: *American Economic Review* 99.3, pp. 909–936.

Kozakiewicz, Marta (2020). "Belief-Based Utility and Signal Interpretation". Working Paper.

Lerner, Jennifer S, Ye Li, Piercarlo Valdesolo, and Karim S Kassam (2015). "Emotion and decision making". In: *The Annual Review of Psychology* 66.

Macera, Rosario (2014). "Dynamic beliefs". In: *Games and Economic Behavior* 87, pp. 1–18.

Malmendier, Ulrike and Geoffrey Tate (2005). "Does overconfidence affect corporate investment? CEO overconfidence measures revisited". In: *European Financial Management* 11.5, pp. 649–659.

— (2008). "Who makes acquisitions? CEO overconfidence and the market's reaction". In: *Journal of Financial Economics* 89.1, pp. 20–43.

McCarty, Nolan, Keith T Poole, and Howard Rosenthal (2016). *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

Mezulis, Amy H, Lyn Y Abramson, Janet S Hyde, and Benjamin L Hankin (2004). "Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias." In: *Psychological Bulletin* 130.5, p. 711.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat (2022). "Managing self-confidence: Theory and experimental evidence". In: *Management Science* 68.11, pp. 7793–7817.

Niederle, Muriel and Lise Vesterlund (2007). "Do women shy away from competition? Do men compete too much?" In: *The Quarterly Journal of Economics* 122.3, pp. 1067–1101.

Ortoleva, Pietro and Erik Snowberg (2015). "Overconfidence in political behavior". In: *American Economic Review* 105.2, pp. 504–35.

Pekrun, Reinhard, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P Perry (2011). "Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ)". In: *Contemporary Educational Psychology* 36.1, pp. 36–48.

Schlag, Karl H, James Tremewan, and Joël J Van der Weele (2015). "A penny for your thoughts: a survey of methods for eliciting beliefs". In: *Experimental Economics* 18.3, pp. 457–490.

Schwardmann, Peter and Joel Van der Weele (2019). "Deception and self-deception". In: *Nature Human Behaviour* 3.10, pp. 1055–1061.

Schwartzstein, Joshua (2014). "Selective attention and learning". In: *Journal of the European Economic Association* 12.6, pp. 1423–1452.

Shepperd, James A, Judith A Ouellette, and Julie K Fernandez (1996). "Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback." In: *Journal of Personality and Social Psychology* 70.4, p. 844.

Spielberger, Charles D (1983). "State-Trait Anxiety Inventory for Adults". In: *PsycTESTS Dataset*.

Sweeny, Kate, Patrick J Carroll, and James A Shepperd (2006). "Is optimism always best? Future outlooks and preparedness". In: *Current Directions in Psychological Science* 15.6, pp. 302–306.

Sweeny, Kate and Angelica Falkenstein (2015). "Is waiting the hardest part? Comparing the emotional experiences of awaiting and receiving bad news". In: *Personality and Social Psychology Bulletin* 41.11, pp. 1551–1559.

Sweeny, Kate and Zlatan Krizan (2013). "Sobering up: A quantitative review of temporal declines in expectations." In: *Psychological Bulletin* 139.3, p. 702.

Uusberg, Andero, Jamie L Taxer, Jennifer Yih, Helen Uusberg, and James J Gross (2019). "Reappraising reappraisal". In: *Emotion Review* 11.4, pp. 267–282.

Van den Steen, Eric (2004). "Rational overoptimism (and other biases)". In: *American Economic Review* 94.4, pp. 1141–1151.

Van Dijk, Wilco W, Marcel Zeelenberg, and Joop Van der Pligt (2003). "Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment". In: *Journal of Economic Psychology* 24.4, pp. 505–516.

Zimmermann, Florian (2020). "The dynamics of motivated beliefs". In: *American Economic Review* 110.2, pp. 337–61.