Hope for the best, prepare for the worst: signal anticipation and ex-ante belief manipulation*

Marta Kozakiewicz†

June 24, 2025

Abstract

In this paper, 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. Furthermore, I examine a previously unexplored link between gain-loss attitudes and overconfidence, and confirm it in the data.

*The author gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (project A01). Financial support from the German Research Foundation (DFG project 462020252) is gratefully acknowledged.

†Frankfurt School of Finance and Management; email: m.kozakiewicz@fs.de

1 Introduction

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. 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, hence also the consumption-savings decisions. A large body of work has shown that people tend to overestimate their abilities and chances of success. While many papers document overconfident beliefs and their consequences, there are still open questions about the way overconfidence arises and why it persists.[1] In this paper, I argue that the answer lies in the factors driving 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, yet the magnitude and the direction of the effect differ across studies (Benjamin, 2019).

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, 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 (Bracha and Brown, 2012; Brunnermeier and Parker, 2005; Caplin and Leahy, 2019), others describe conditions under which the agent adopts pessimistic beliefs (Gollier and Muermann, 2010; Macera, 2014). The theoretical framework presented in this paper builds upon the latter. I assume that the agent's utility is reference-dependent (Kőszegi and Rabin, 2006), and the current belief level serves as a reference point.[2]

[1] For a comprehensive review of the literature, I refer the reader to Huffman et al. (2022).

[2] The idea of reference-dependent utility from beliefs was first proposed by Kőszegi and Rabin (2009). However, their approach is conceptually different, as beliefs are not subject to a choice but are rationally

The intuition behind the model presented in this paper is that people can "prepare themselves" for the arrival of new information by adopting pessimistic beliefs. The prior is not fixed, like in the classical approach, 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, e.g., Carroll et al., 2006; Shepperd et al., 1996; Sweeny et al., 2006; Sweeny and Krizan, 2013), none of the studies look at a shift in beliefs before a partial information revelation.[3] Why would people brace themselves before 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. 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, when an agent expects to receive a noisy signal, he chooses beliefs that are lower (more pessimistic) than the beliefs he would choose 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

formed based on credible plans of future behavior. In the rational-expectations models, an agent derives utility from beliefs that are consistent with his actions in equilibrium (Caplin and Leahy, 2001; Kőszegi, 2010; Kőszegi and Rabin, 2009).

[3] 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.

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.[4]

In order to test the model predictions, I designed a simple experiment. First, subjects solved an IQ test that allowed me to measure their cognitive ability. 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. I delayed the full information revelation to minimize confounding factors (see footnote 7) 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* 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 an independent measure of loss aversion.[5,6]

The experiment was conducted in the summer of 2022 at the University of Bonn. I collected data from 234 participants, mostly university students. Subjects revealed the mean belief above the 6^{th} decile of the distribution—significantly higher than the average actual position. To test the first prediction of the model, I compare the mean beliefs in the two conditions. Unfortunately, I cannot confirm the first hypothesis: the average treatment effect is negative but not significant at any acceptable level. The results suggest that the reference-dependence component of the utility function, albeit present, is not strong enough to counteract the incentive to maintain optimistic beliefs.

[4]Here, I refer to the absolute bias: the absolute difference between one's beliefs and their ability.

[5]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.

[6]The measure was obtained using 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 concerning the utility they would experience before receiving a signal. Participants who reported lower anticipatory utility are considered to be more loss averse.

In order to test the second prediction, I examine the relationship between participants' beliefs and their loss attitudes. 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 less pronounced for participants with higher ability, that is, 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).

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: the way loss aversion drives subjects' beliefs in the two conditions. Moreover, the evidence on over- and underconfidence further support the theory.

This paper contributes to the literature on belief formation when the decision-maker derives utility from his beliefs. 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.[7] This includes an experimental study by Van Dijk et al. (2003), which provides evidence on lowering one's beliefs before learning a test score.[8] 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.[9] Belief choice in a different setting was also studied in Engelmann et al. (2019). The authors provide evidence on "wishful thinking" (Caplin and Leahy, 2019). In contrast to my work, they study the formation of beliefs about outcome that is not relevant to agent's self-esteem. Other experimental literature focused so far on documenting deviations from the rational update (Coutts, 2019; Drobner and Goerg, 2024; Eil and Rao, 2011; Ertac, 2011; Kozakiewicz, 2020; 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.

[7] 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 postponing the full information revelation, the strategy becomes less salient, so one can focus on the trade-off described in the model.

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

[9] 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.

More broadly, my work contributes to the literature on motivated reasoning (see Bénabou and Tirole, 2016). I add to this literature by providing evidence on the functional form of belief-based utility, which can inform the theory and advance development of models describing phenomena such as information avoidance or belief polarization. Moreover, this is the first paper to establish a link between gain-loss attitudes, ability, and confidence. The results provide the first evidence on *who* is more likely to become overconfident (low-ability and non-loss-averse agents), complementing the literature that investigates the origins of overconfidence (see, for example, Burks et al., 2013; Huffman et al., 2022; Schwardmann and Van der Weele, 2019). The paper proceeds as follows. In the next section, I describe the model. The experimental design is outlined in Section 3. In Section 4, 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 5. The last section concludes.

2 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 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 $u(p_t) = p_t u_H + (1 - p_t)u_L$.[10] 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.[11] Moreover, the agent knows that he will receive a signal in Period 1

[10] 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.

[11] 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.

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:

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

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 $P(s=L|p_0) = 1 - P(s=H|p_0)$. The maximization is subject to the following inequality constraints: $\tilde{p}_0 \leq 1$ and $\tilde{p}_0 \geq 0$. The model is very similar to Gollier and Muermann (2010), as the agent chooses subjective probabilities facing a trade-off between ex-ante 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. For simplicity, I assume that the agent assigns equal weight to the current utility and the utility experienced after a signal (the discount factor β is equal to one and omitted in the formula). 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.

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; Caplin and Leahy, 2019; Gollier and Muermann, 2010; Macera, 2014). 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.[12] These assumptions ensure that belief formation is not entirely detached from reality, but partly driven by a more objective, rational process. One can also view this as a modeling technique, conceptually similar to a dual-self model (Fudenberg and Levine, 2006). In this case, it is used to describe an internal, subconscious process of coming to the belief about one's ability.

In the internal solution, the first-order condition gives us the following:

$$\tilde{p_0}^* \;=\; p_0 \;+\; \frac{1}{\gamma}\Big(1 - P(s = H|p_0)\;\eta - P(s = L|p_0)\;\lambda\eta\Big)(u_H - u_L). \tag{2}$$

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 enjoy optimistic belief in Period 0.

It is easy to show that, for loss-averse individuals ($\lambda > 1$), the chosen belief is always lower than the unmanipulated belief p_0. For non-loss-averse agents ($0 < \lambda < 1$), it is higher. In other words, loss-averse subjects will manipulate their beliefs *downwards* and non-loss-averse individuals *upwards* from the rational belief. The loss-neutral agents ($\lambda = 1$) would choose a belief equal to the unmanipulated belief: $\tilde{p_0}^* = p_0$. Parameter restrictions, which ensure that $p^* \in [0, 1]$, are discussed in Appendix B.1. An agent might hit a boundary if he is 1) sufficiently loss-averse with low p_0, or 2) sufficiently gain-seeking with high p_0. I will show that the results are indeed stronger, when I exclude these subjects from the analysis.

2.1 No information about a signal

How one can describe an agent who does not know that he will receive a signal? I assume that receiving a signal is not a zero-probability event – the agent assigns a probability ϵ

[12] 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.

to it.[13] He chooses $\hat{p}_0$ to maximize:

$$U_0 = u(\tilde{p}_0) + \underbrace{\epsilon\ P(s=H|p_0)\left[u(p_1^H)\ +\ \mu\Big(u(p_1^H) - u(\tilde{p}_0)\Big)\right]}_{\text{the utility in Period 1}} + \\ + \underbrace{\epsilon\ P(s=L|p_0)\left[u(p_1^L)\ +\ \mu\Big(u(p_1^L) - u(\tilde{p}_0)\Big)\right]}_{\text{the utility in Period 1}} - \frac{\gamma}{2}(\tilde{p}_0 - p_0)^2, \tag{3}$$

subject to the following inequality constraints: $\tilde{p}_0 \leq 1$ and $\tilde{p}_0 \geq 0$. For $\epsilon = 1$, the above equation reduces to (1), that is, the problem of an agent expecting a signal. For $\epsilon \in (0,1)$, the first-order condition gives us:

$$\hat{p}_0^{\,*} \ =\ p_0\ +\ \frac{1}{\gamma}\Big(1 - \epsilon P(s=H|p_0)\eta - \epsilon P(s=L|p_0)\lambda\eta\Big)(u_H - u_L). \tag{4}$$

For $\lambda > 0$ and $\eta = 1$, we have: $\tilde{p}_0^{\,*} \leq \hat{p}_0^{\,*}$. The equality holds only on the boundary (see Appendix). In the interior solution, the optimal prior chosen when expecting a signal is always lower than the prior chosen not knowing about a signal. Intuitively, when the agent does not expect a signal, he does not take into account the prospect of being disappointed or elated by a signal. The incentives to lower beliefs 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.

Proposition 1

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

Proposition 1, together with propositions presented in the following sections, guide the experimental design and underlie the testable predictions presented in Section 4. All proofs, as well as the derivations of parameter restrictions, are delegated to Appendix B.1.

[13] 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.

It is important to note that the parameter restrictions ensuring the interior solution in the case of no signal depend on the value of ϵ. For ϵ close to 1, the conditions are similar to the conditions in the case of expecting a signal: an agent might hit a boundary if he is 1) sufficiently loss-averse with low p_0, or 2) sufficiently gain-seeking with high p_0. For ϵ close to 0, however, no agent will hit the lower bound, but any agent with sufficiently high p_0 will hit the upper bound. See Appendix B.1 for details.

2.2 Comparative statics: loss aversion

The optimal solution (2) depends on the loss-aversion parameter:

$$\frac{\partial \tilde{p}_0^{\,*}}{\partial \lambda} = -\frac{1}{\gamma}\ \eta\ P(s = L|p_0)(u_H - u_L) < 0. \tag{5}$$

In the interior solution, 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 proposition:

Proposition 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 interior solution depends on λ as long as ϵ is larger than zero:

$$\frac{\partial \hat{p}_0^{\,*}}{\partial \lambda} = -\frac{1}{\gamma}\ \epsilon\ \eta\ P(s = L|p_0)(u_H - u_L) < 0. \tag{6}$$

If the ϵ is equal to zero, there will be no effect of loss-aversion on the optimal prior. Comparing the two conditions, one can see that:

$$\frac{\partial \tilde{p}_0^{\,*}}{\partial \lambda} < \frac{\partial \hat{p}_0^{\,*}}{\partial \lambda}. \tag{7}$$

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.

Proposition 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.

2.3 Comparative statics: loss aversion and unmanipulated beliefs

The probability of receiving a high signal, $P(s = H|p_0)$, depends on a signal structure and the unmanipulated belief p_0. I 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, he 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 λ and p_0 gives us the following:

$$\frac{\partial \partial \tilde{p}_0^{\,*}}{\partial \lambda \partial p_0} = \frac{1}{\gamma}\eta(2c - 1)(u_H - u_L) > 0. \tag{8}$$

Since $c > \frac{1}{2}$ and $u_H > u_L$, the derivative has a positive sign. Thus, an increase in the unmanipulated belief p_0 mitigates the negative effect of loss aversion. I conclude that:

Proposition 3.

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

2.4 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. One can obtain the formula by moving p_0 to the other side of equation (2):

$$\tilde{p}_0^{\,*} \; - \; p_0 \; = \; \frac{1}{\gamma}\Big(1 - P(s = H|p_0)\ \eta - P(s = L|p_0)\ \lambda\eta\Big)(u_H - u_L). \tag{9}$$

If p_0 perfectly captures the state ($p_0 = 0$ if the agent's ability is low, $p_0 = 1$ if high), one can draw a link between the agent's characteristics and his bias. A positive (negative) bias is indicative of overconfidence (underconfidence). Assuming $\eta = 1$, $\gamma > 0$, and $u_H > u_L$, the right hand side of (11) is positive for $\lambda < 1$ and negative for $\lambda > 1$. Secondly, a high-ability, gain-seeking agent chooses the boundary solution $\tilde{p_0}^* = 1$. He cannot adopt beliefs higher than $p_0 = 1$, so he cannot be overconfident. Similarly, a low-ability, loss-averse agent cannot be underconfident. In these cases, and in the case of $\lambda = 1$, the agent is unbiased. Taken together:

Proposition 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.
iii) an agent is unbiased if and only if one of the following is true: he is loss-neutral, or he is high-ability and loss-averse, or he is low-ability and gain-seeking.

Moreover, the extent of 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 2.3:

$$\tilde{p_0}^* - p_0 = \frac{1}{\gamma}(p_0 - 2cp_0 + c)(1 - \lambda)(u_H - u_L), \tag{10}$$

where c is signal precision, $c > 0.5$. The middle term is the probability of receiving a "bad" signal, $P(s = L|p_0)$. For low-ability agents ($p_0 = 0$), this term reduces to $(p_0 - 2cp_0 + c) = c$. For high-ability individuals ($p_0 = 1$), this term becomes $(1 - c)$. It is important to note that, since $c > 0.5$ (signals are informative), we have $c > 1 - c$. Thus, the absolute bias, understood as the distance between the manipulated belief and the true state, will be larger for low-ability than for high-ability agents. Second, because the LHS of (10) is positive (negative) for low-ability (high-ability) agents, the absolute bias will be decreasing (increasing) in the loss aversion parameter. Lastly, the absolute bias will be more responsive to changes in λ for low-ability agents (the loss aversion

parameter is multiplied by $c > 1 - c$).

Proposition B.2

i) The absolute bias is larger for low-ability than for high-ability agents.

ii) The absolute bias is decreasing in the loss aversion parameter λ for low-ability agents, and increasing in λ for high-ability individuals.

iii) λ has a stronger impact on the absolute bias for low-ability agents.

Another interesting implication of (10) is that the absolute bias is increasing in c for low-ability and decreasing in c for high-ability agents.[14] The precision of a signal affects agents differently, depending on their type. High-ability, underconfident agents become *more* accurate, whereas low-ability, overconfident individuals *less* accurate in their assessments.[15] For low-ability agents, the intuition is the following: when 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 high-ability 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.

How do the predictions differ for participants who do not expect a signal? In this case, the predictions depend on the value of ϵ. For ϵ close to one, they are similar to Proposition B.1. For lower ϵ, the value of λ distinguishing over- and under-confident subjects increases, and at ϵ close to zero it is so high that virtually *all* subjects are predicted to be overconfident (see Appendix B.4). One can easily see it by solving (3) with $\epsilon = 0$: the optimal belief $\tilde{p}_0^*$ is always higher than p_0 indicating a positive bias. Proposition B.2 remains the same when not expecting a signal, however, the effects will be attenuated, and the extent of attenuation will depend on the value of ϵ.

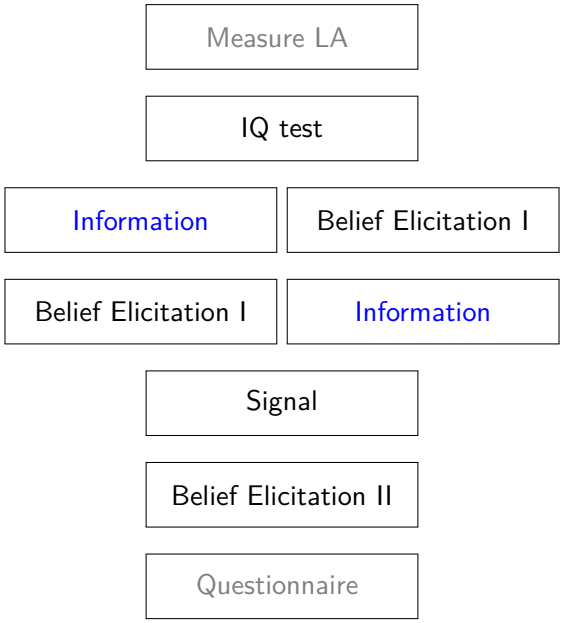
[14] The experiment was not designed to test this prediction, however, one could easily modify my design to check how the agent's bias changes when one increases the signal precision.

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

3 Experimental Design

In this section, I describe the experimental design. The outline of the experiment is presented in Figure 1. The experiment consisted of several parts, which were the same in the Treatment and in the Control condition.[16] The two conditions differed only with respect to the timing of information provision (marked in blue in Figure 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* I elicited their beliefs.

Figure 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 will be calculated 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

[16]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).

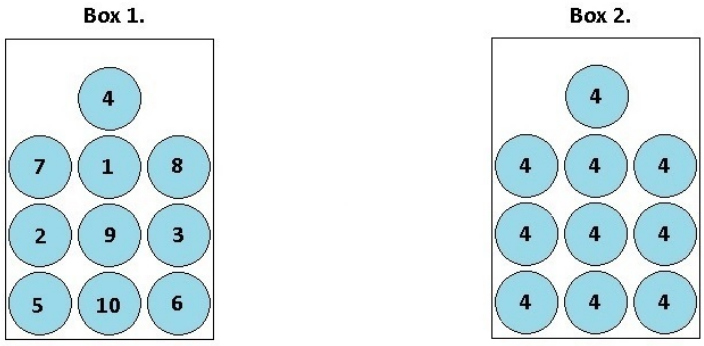


Figure 2: Boxes of a person whose rank was 4 (the "Signal" condition).

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. Next, subjects 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. Additionally, I introduced a procedure to obtain a measure of individual loss aversion. It was implemented at the very beginning of the experiment, before the IQ test. I describe it in detail in Section 3.2.

3.1 Signal Structure and Belief Elicitation

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 participants with the lowest test 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

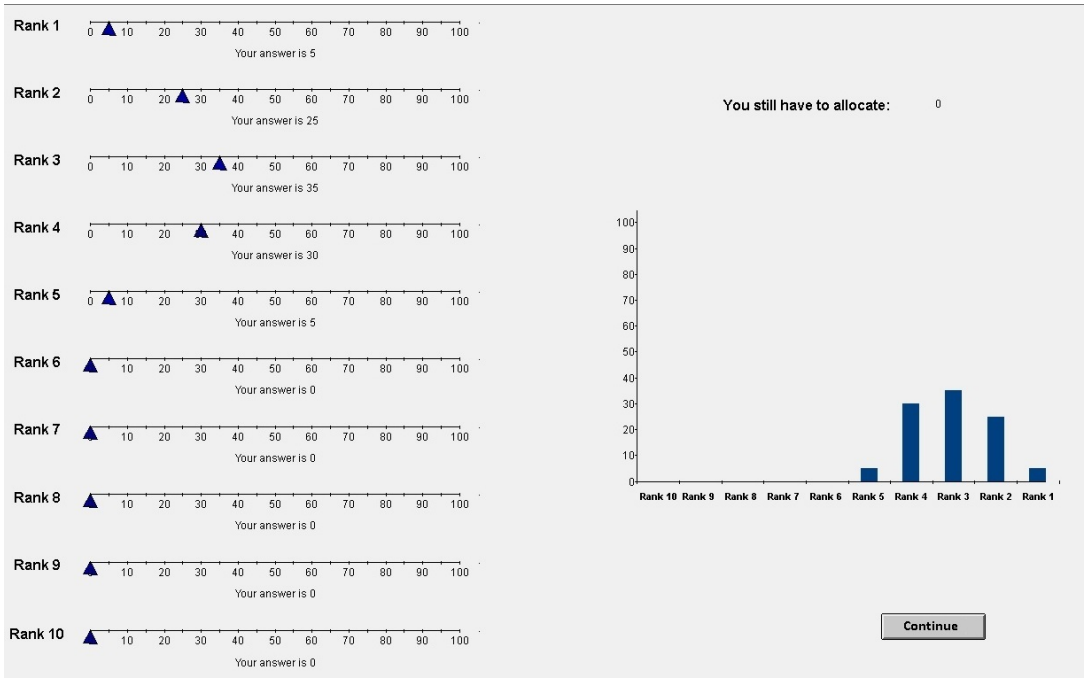


Figure 3: The screen-shot of the interface used by subjects in belief elicitation.

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 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 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 structure, we explained to the subjects the belief elicitation task. Participants were asked to report their beliefs by allocating 100 points among 10 ranks using a computer interface (see Figure 3). They were allocating points by dragging blue arrows and were able to correct their choices if they wished. The allocation immediately appeared on the graph to the right. The text above the graph informed subjects how many points they still have to allocate before they can proceed to the next task.

To incentivize truthful reporting, I used the Binarized Scoring Rule (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 used the same instructions for the belief elicitation as in the Treatment condition. The only difference is that in the Control condition, participants were not given information about the upcoming signal. This information, together with the explanation of the signal structure, was given to the subjects immediately after they finished the elicitation task.

3.2 Measuring Loss Aversion

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:

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 answered 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:

$$\begin{aligned} U_Q &= P(s=H|p_0)\ \left[u(p_1^H)+\eta\Big(u(p_1^H)-u(p_0)\Big)\right]+ \\ &\quad + P(s=L|p_0)\ \left[u(p_1^L)+\eta\lambda\Big(u(p_1^L)-u(p_0)\Big)\right], \end{aligned} \tag{11}$$

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:

$$\begin{aligned} U_Q \;&=\; 0.5\left\{0.7u_H+0.3u_L+\eta\left[0.7u_H+0.3u_L-\Big(0.5u_H+0.5u_L\Big)\right]\right\}+ \\ &+\; 0.5\left\{0.3u_H+0.7u_L+\eta\lambda\left[0.3u_H+0.7u_L-\Big(0.5u_H+0.5u_L\Big)\right]\right\}= \quad (12) \\ &=\; 0.5\;(u_H+u_L)\;+\;0.1\;\eta\;(1-\lambda)\;\Big(u_H-u_L\Big)\implies \lambda=a\;U_Q+b. \end{aligned}$$

Where $a=\frac{-10}{\eta(u_H-u_L)}$ and $b=1+\frac{5(u_H+u_L)}{\eta(u_H-u_L)}$. If we assume that u_H and u_L are the same for all participants, the differences in the anticipatory utility U_Q reflect the differences in the loss aversion parameter λ. The two hypothetical 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.[17]

I also introduced a second, incentivised measure of loss aversion that was based on the willingness to pay for a signal conditional on its realization.[18] Unfortunately, the attempt to obtain this measure 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.[19] The description of the measure, the collected data, and the results based on it, as well as the theory behind it, are gathered in Appendices D, E, and F.

[17] 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 the model.

[18] The procedure was introduced only after the first belief elicitation, therefore, it does not affect the results presented in the following sections. 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.

[19] Still, 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 (see Appendix D and E).

4 Testable Predictions

The main outcome variable is the optimal prior $\tilde{p}_0^{\ *}$ reported in Belief Elicitation I. By comparing the average belief in the Treatment and in the Control condition, one can shed light on Proposition 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 also predicts 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 (Proposition 2.1). In the Control condition, this correlation should be weakly negative due to the discount factor $\epsilon \in (0, 1)$. However, the negative effect should be weaker than in the Treatment condition (Proposition 2.2). To test these predictions, I look at the following regression:

$$Prior_i = \alpha_0 \ + \ \alpha_1 \, Treatment_i \ + \ \alpha_2 \, \lambda_i \ + \ \alpha_3 \, Treatment_i \times \lambda_i \ + \ \epsilon_i, \qquad (13)$$

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 ($Treatment_i = 0$). The correlation between beliefs and gain-loss attitudes in the Treatment condition is captured by $\alpha_2 + \alpha_3$. The coefficient α_3 informs us if the increase in the loss aversion parameter has a larger negative effect in the Treatment compared to the Control condition. The predictions are summarized in Hypothesis 2.

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* (13) *is negative.*

ii) In the Control condition, the correlation is weakly negative. The coefficient α_2 *in* (13) *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* (13) *is negative.*

In order to test Prediction 3, I use a specification that allows for examining a joint effect of the loss aversion parameter λ and unmanipulated beliefs p_0. Note that I 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:

$$Prior_i = \beta_0 \ + \ \beta_1 \ Ability_i \ + \ \beta_2 \ \lambda_i \ \ + \ \beta_3 \ Ability_i \times \lambda_i \ + \ \epsilon_i, \tag{14}$$

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 3, this effect should be positive. Considering that an increase in λ has a negative effect on prior beliefs, an increase in participant's ability lessens the negative effect of loss aversion. Lastly, I examine the joint effect of the treatment, loss aversion, and unmanipulated beliefs. I look at the following regression:

$$\begin{aligned} Prior_i = \ & \gamma_0 \ + \ \gamma_1 \ Treatment_i \ + \ \gamma_2 \ \lambda_i \ + \ \gamma_3 \ Ability_i \ + \ \gamma_4 \ Ability_i \times \lambda_i \ + \\ & + \ \gamma_5 \ Treatment_i \times \lambda_i \ + \ \gamma_6 \ Treatment_i \times Ability_i \ + \ \epsilon_i. \end{aligned} \tag{15}$$

The coefficient γ_4 should be positive, providing additional evidence for Prediction 3.

Hypothesis 3

The effect of loss aversion on prior beliefs is less negative for subjects with higher ability. The coefficient β_3 *in* (14)*, as well as* γ_4 *in* (15)*, is positive.*

5 Data Analysis

In this section, I present the results of the data analysis. 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 are similar, albeit noisier, if I include those subjects in the analysis (see Appendix C).

5.1 Raw Data

The distribution of participants' IQ test scores is presented in Figure 4. The average test score was 5.17, with a standard deviation of 3.42. Table 10 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 1).[20]

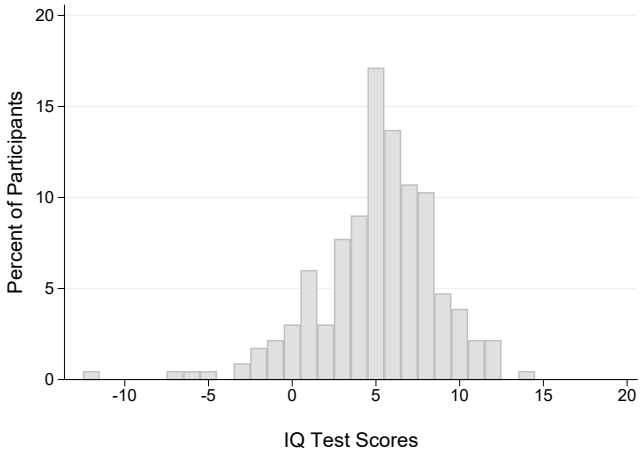


Figure 4: Distribution of the IQ test scores.

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,

[20]Due to the differences in the instructions, I could not randomly assign the treatment status to participants *within* a session. I 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 I collected: loss aversion, cognitive reappraisal, and suppression.

Table 1: Differences between participants in the two conditions.

	Treatment	**Control**		Diff < 0	Diff $\neq 0$	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 remaining subjects reported right- or left-skewed distributions.[21] 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).

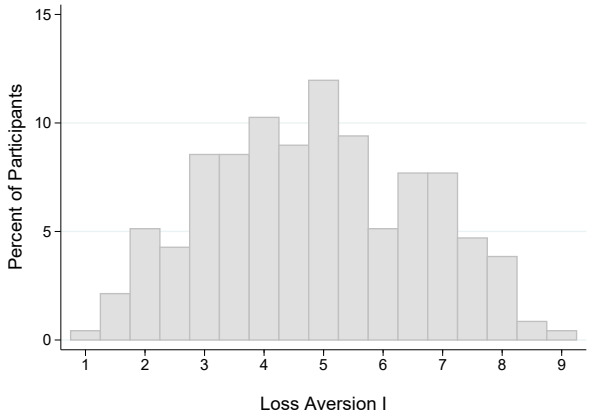
Participants in the sample tend to be *overconfident*, revealing belief distributions with means that were, on average, 1 rank lower than their true rank.[22] Note that, in this set-up, a lower rank corresponds to a better performance. Similarly, a lower mean belief (expressed in terms of rank) indicates 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.

The measure of loss aversion is defined as the average response to the two hypothetical questions. Its distribution is presented in Figure 5. 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 Table). At the same time, the measure of loss aversion correlates with beliefs as predicted by the model—I explore this relationship in the following sections.

[21] 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.

[22] 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 G.

Figure 5: Distribution of loss aversion in the entire sample.



5.2 Testing Model Predictions

5.2.1 Hypothesis 1

First, 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.[23] It takes values from 1 to 10, with higher values indicating beliefs in obtaining a higher test result.[24] The dependent variable is regressed on a treatment dummy, and the results are gathered in Table 2. The first two columns contain the estimates based on the sample of all participants. The coefficient at the "Treatment" variable is not significant at any acceptable level. In the last columns in Table 2, I present the estimates based on the restricted sample. The treatment effect is negative, as predicted by the theory, but far from significant: p-values of two-sided t-test =0.519 and =0.875 (with and without controlling for subject's ability).

Result 1

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

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

[24]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 2: The effect of treatment on mean beliefs.

	All subjects		Restricted sample	
Treatment	-0.141	-0.026	-0.149	-0.035
	(0.225)	(0.218)	(0.230)	(0.224)
Ability		0.178***		0.175***
		(0.040)		(0.042)
N	227	227	218	218

The dependent variable is the mean belief revealed in Belief Elicitation I. "All subjects" refers to the whole sample. "Restricted sample" refers to subjects for whom the effect is expected. Standard errors in parentheses. All specifications include a constant (omitted for clarity).
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5.2.2 Hypothesis 2

Before I test Hypothesis 2, I plot the averages of subjects' beliefs for groups with different loss aversion parameters (see Figure 6). 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 the red line corresponds to the regression fitted on the data from the Treatment and the Control condition, respectively.

Figure 6: The mean belief depending on loss aversion in the two conditions.

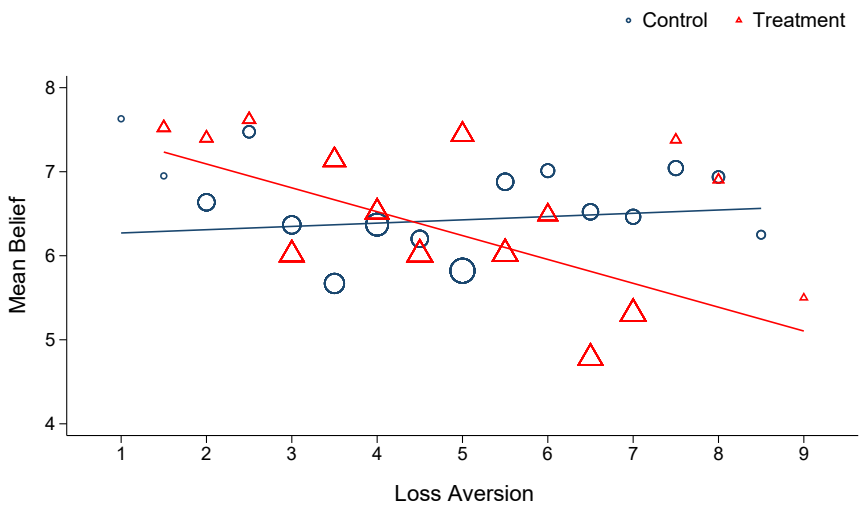


Table 3: The effect of treatment and loss aversion on mean beliefs.

	All subjects		Restricted sample (R)	
	(1)	(2)	(1)	(2)
Treatment	1.038	1.049*	1.428**	1.354**
	(0.657)	(0.632)	(0.687)	(0.662)
Loss Aversion	-0.021	-0.022	0.039	0.002
	(0.085)	(0.082)	(0.091)	(0.088)
Treatment × Loss Aversion	-0.238*	-0.218*	-0.323**	-0.283**
	(0.127)	(0.123)	(0.134)	(0.130)
Ability		0.173***		0.175***
		(0.040)		(0.042)
N	227	227	218	218

The dependent variable is the mean belief revealed in Belief Elicitation I. "Loss Aversion" takes values from 1 to 9 and denotes the average response to the hypothetical questions. "Ability" takes values from 1 to 10 and denotes the position in the IQ test distribution Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

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.

The observation is confirmed in a regression analysis presented in Table 3 The first two columns contain the estimates based on the entire sample, and the last columns 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. 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 is not significantly dif-

ferent 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 0.03, the interaction effect is significant at the 5% level. Based on these results, I confirm the second prediction of the model.

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 individuals lower their beliefs more in the Treatment condition.

5.2.3 Hypothesis 3

In order to test Hypothesis 3, I run the regression specified in (14). The results are presented in Table 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. I also 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 loss aversion on the optimal belief. The result prevails when I add the remaining two-way interactions.

Result 3

The effect of loss aversion on the mean belief is less negative for participants with higher ability. The coefficient β_3 in (14) *is positive.*

Table 4: The effect of treatment, loss aversion, and unmanipulated beliefs.

	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Treatment	0.011	0.507	0.041	0.658
	(0.215)	(0.767)	(0.220)	(0.766)
Loss Aversion	-0.403***	-0.297**	-0.528***	-0.367**
	(0.136)	(0.150)	(0.157)	(0.185)
Ability	-0.083	-0.107	-0.173	-0.212
	(0.118)	(0.123)	(0.134)	(0.137)
Ability × LA	0.053**	0.050**	0.073***	0.068***
	(0.023)	(0.023)	(0.026)	(0.026)
Treatment × LA		-0.194		-0.270**
		(0.122)		(0.128)
Treatment × Ability		0.083		0.127
		(0.079)		(0.082)
N	227	227	218	218

"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 test scores.
All specifications include a constant (omitted for clarity).
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5.3 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 2.4.

5.3.1 Prediction B.1

In order to test Prediction B.1, I classify subjects based on relevant characteristics. A subject is classified as "high-ability" if his rank was below or equal to 5. Participants whose ranks were above 5 are classified as "low-ability". A subject is "loss-averse" ("non-loss-averse") if the average of his responses to the hypothetical questions was above (below) 5. I classified a person as overconfident (underconfident) if his mean belief was more than 1 rank higher (lower) than his actual rank. Subjects with the

Table 5: Frequency of types among low-ability subjects.

	Non loss-averse	Neutral	Loss-averse
Underconfident	3	1	5
Unbiased	4	1	8
Overconfident	51	6	38

Table 6: Frequency of types among high-ability subjects.

	Non loss-averse	Neutral	Loss-averse
Underconfident	29	12	24
Unbiased	13	4	9
Overconfident	11	2	6

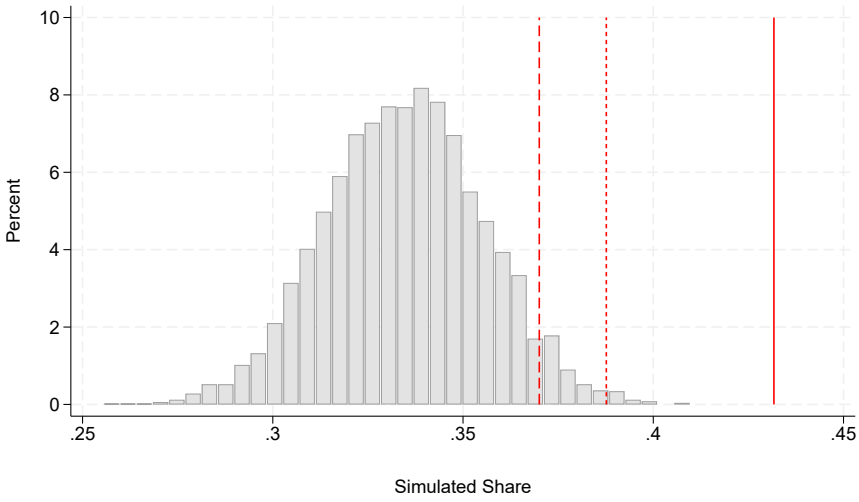
mean belief within ± 0.5 rank from the actual performance are classified as unbiased. In Table 5, I show the frequency of each type in the group of low-ability individuals. Similar frequencies for high-ability subjects are presented in Table 6. One can notice that the types are not distributed evenly: for example, in the group of low-ability, non-loss-averse subjects, 51 out of 58 are overconfident (88%). Due to the presence of people who are underconfident and non-loss-averse (overconfident and loss-averse), the strong version of Proposition B.1 does not hold. Still, I find it worthwhile to test a weaker version of this hypothesis: the types predicted by the model are over-represented in the data.

To this end, I classify a subject as "in line with the model" if his bias was in the direction predicted by his ability and the loss-aversion parameter.[25] The share of participants whose types coincide with the model is 43.2%. However, since the relative performance is bounded, some relations can arise mechanically. In particular, low-ability individuals are more likely to be overconfident, and high-ability subjects—underconfident.[26] To answer the question whether confidence types occur more often in the data than they would mechanically, I look at a distribution of types when beliefs are drawn from the

[25]The classification of subjects from the Control condition depends on the value of ϵ. In the main text, I use the classification based on the assumption of a low ϵ. The results with a different classification—assuming a high value of ϵ—is delegated to Appendix.

[26]However, there should be no mechanical relation between loss aversion and rank or beliefs.

Figure 7: Coefficients in the simulation (gray bars) and in the data (solid red line).



uniform distribution. I simulate it 5000 times, each time calculating the share of types that coincide with the types predicted by the model.

As one can see on Figure 7, the fraction of people whose types are aligned with the model predictions is significantly higher than they would be if beliefs were assigned at random (the dashed red lines indicate the 95^{th} and the 99^{th} percentile of the distribution). The results are similar when we assign at random not only beliefs but also loss-aversion parameters. In Appendix, I present simulation with the loss-aversion parameter drawn from either the uniform distribution or the distribution present in the data. Moreover, I perform the same exercise with the beliefs drawn from distribution centered around the subject's actual rank: instead of the uniform distribution over the entire domain, I restrict the draw to be no further than 2 ranks from the actual performance. The result is robust to these manipulations.

Result B.1

The observed combination of the loss-aversion parameters, abilities, and biases, is significantly more likely to be generated by the mechanism of the model than to be a result of a random process.

5.3.2 Prediction B.2

In order to test the second prediction of the model, I use the following specification:

$$Bias_i = \beta_0 \ + \ \beta_1 \ AbilityLow_i \ + \ \beta_2 \ \lambda_i \ + \ \beta_3 \ AbilityLow_i \times \lambda_i \ + \ \epsilon_i, \tag{16}$$

where $Bias_i$ denotes the absolute bias, $AbilityLow_i$ is an indicator variable taking value 1 if the agent's rank is one of the worst five ranks, and λ_i is the measure of loss aversion. The results are gathered in Table 7. As in the previous sections, I present two sets of results: estimates based on the sample of all participants and estimates based on the restricted sample (with the same exclusion criteria).

First, I regress the dependent variable, the absolute bias, on the indicator variable "Low Ability". The coefficient is positive and highly significant: the average bias of overconfident subjects is 1.491 rank higher than the bias of underconfident participants (96% increase in relative terms). This result confirms the first part of Prediction B.2. As an additional exercise, I simulated Specification (1) from Table 7 using beliefs drawn from the uniform distribution. The coefficient at the 99^{th} percentile equals 0.960, a value

Table 7: The effect of overconfidence and loss aversion on the absolute bias.

	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Low Ability	1.491***	2.266***	1.451***	2.684***
	(0.217)	(0.632)	(0.221)	(0.655)
Loss Aversion		-0.024		0.007
		(0.088)		(0.091)
Low Ability × Loss Aversion		-0.158		-0.259**
		(0.123)		(0.128)
Const.	1.555***	1.670***	1.537***	1.568**
	(0.156)	(0.449)	(0.158)	(0.472)
N	227	227	218	218

The dependent variable is the agent's bias (the absolute value). "Loss Aversion" denotes the average response to the hypothetical questions. "Low Ability" is a dummy indicating an agent with a rank above 5. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

much lower than 1.491 (the actual coefficient). This result confirms that the difference in biases between low-ability and high-ability agents would not arise if beliefs were not subject to motivated reasoning.

In the second and the last column, I gather the results of estimating (16). One can notice that restricting the sample significantly changes the estimates. As expected, 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 I only interpret coefficients based on the restricted sample.

For low-ability subjects, there is a negative and significant effect of loss aversion on the absolute bias. In line with the model, more loss-averse subjects end up with a smaller bias. For high-ability participants, there is no effect. Proposition B.2 iii) provides us with an explanation for why this might be the case. The probability of receiving a "bad" signal is lower than for high-ability subjects. Consequently, the loss component is downweighted and the effect of loss aversion is diminished. While the probability of a "bad" signal is never zero (there is always the noise component), the effect might be hard to detect in a small sample.

Result B.2

i) The absolute bias is larger for low-ability than for high-ability agents.
ii) The absolute bias is decreasing in the loss aversion parameter λ for low-ability subjects, but it is not increasing for high-ability individuals.
iii) λ has a stronger impact on the absolute bias of low-ability participants.

Taken together, the results presented in Sections 5.2 and 5.3 speak in favor of the theory.

6 Conclusions

In this paper, I present an experimental test of a model of belief choice when the utility from beliefs is reference-dependent. 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 belief-based utility. The experimental data show that the gain-loss attitudes drive the process of belief formation. Moreover, the aversion to losses has important implications for the agent's bias. Subjects who overestimate their ability are more likely be low-ability and non-loss-averse—a result that brings us one step closer to understanding the sources of overconfidence. Lastly, it is worth emphasizing 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.

References

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.

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: *Review of Economic Studies* 80.3, pp. 949–983.

Caplin, Andrew and John Leahy (2001). "Psychological expected utility theory and anticipatory feelings". In: *The Quarterly Journal of Economics* 116.1, pp. 55–79.

— (2019). *Wishful thinking*. Tech. rep. National Bureau of Economic Research.

Carroll, Patrick, Kate Sweeny, and James A Shepperd (2006). "Forsaking optimism". In: *Review of general psychology* 10.1, pp. 56–73.

Coutts, Alexander (2019). "Good news and bad news are still news: Experimental evidence on belief updating". In: *Experimental Economics* 22.2, pp. 369–395.

Drobner, Christoph (2022). "Motivated beliefs and anticipation of uncertainty resolution". In: *American Economic Review: Insights* 4.1, pp. 89–105.

Drobner, Christoph and Sebastian J Goerg (2024). "Motivated belief updating and rationalization of information". In: *Management Science*.

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". In.

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.

Fudenberg, Drew and David K Levine (2006). "A dual-self model of impulse control". In: *American economic review* 96.5, pp. 1449–1476.

Goette, Lorenz, Thomas Graeber, Alexandre Kellogg, and Charles Sprenger (2019). "Heterogeneity of gain-loss attitudes and expectations-based reference points". In: *Available at SSRN 3589906*.

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.

Hossain, Tanjim and Ryo Okui (2013). "The binarized scoring rule". In: *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 (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.

Macera, Rosario (2014). "Dynamic beliefs". In: *Games and Economic Behavior* 87, pp. 1–18.

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.

Schwardmann, Peter and Joel Van der Weele (2019). "Deception and self-deception". In: *Nature human behaviour* 3.10, pp. 1055–1061.

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.

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.

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.

A Additional Results

A.1 Results based on the indicator variable

In this section, I test Hypotheses 1-3 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 I used but also the median average response. Based on this definition, there were 90 loss-averse subjects in the 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 8, 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.

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. In Table 8, I present

Figure 8: Beliefs of loss-averse and non-loss-averse subjects in the two conditions.

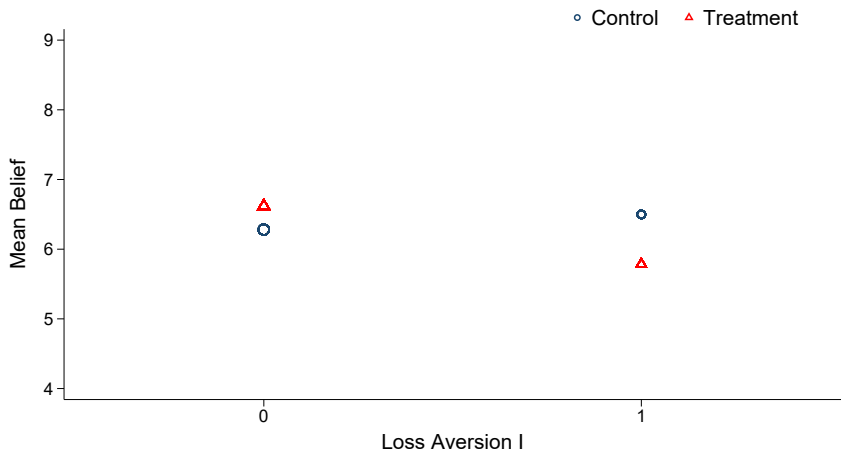


Table 8: The effect of treatment and loss aversion on mean beliefs.

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

the estimates of the same regressions as in Table 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, I run the regressions specified in (14) and (15) using the indicator variable. The results are gathered in Table 9—they correspond to the results presented in Table 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.

Table 9: The effect of treatment, loss aversion, and unmanipulated beliefs.

	All subjects		Restricted sample (R)	
	(1)	(2)	(1)	(2)
Treatment	0.016	0.072	-0.128	-0.519
	(0.218)	(0.513)	(0.230)	(0.566)
Loss Aversion	-0.959**	-0.205	-1.245*	-0.089
	(0.480)	(0.552)	(0.662)	(0.737)
Ability	0.121**	0.097	0.116*	0.041
	(0.052)	(0.061)	(0.060)	(0.074)
Ability × Loss Aversion	0.132	0.099	0.176	0.119
	(0.081)	(0.081)	(0.107)	(0.106)
Treatment × Loss Aversion		-1.126**		-1.551***
		(0.442)		(0.480)
Treatment × Ability		0.072		0.175*
		(0.080)		(0.098)
N	227	227	197	197

"Loss Aversion" is a dummy based on loss aversion measure (1 if average response to the hypothetical questions was above 5). "Ability" denotes position in the IQ test distribution (taking values from 1 to 10, with 10 denoting subjects with the highest scores). All specifications include a constant (omitted for clarity). "All" refers to the sample of all subjects and "R" to the restricted sample.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B Theory

B.1 Parameter restrictions

Expected Signals

Let μ_0 be the Kuhn-Tucker multiplier on the constraint $-\tilde{p}_0 \leq 0$ and μ_1 be the multiplier on the constraint $\tilde{p}_0 \leq 1$. The Lagrangian is given by

$$\mathcal{L} = U(\tilde{p}_0) + \mu_0(\tilde{p}_0) - \mu_1(\tilde{p}_0 - 1)$$

The optimality condition is $\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0} + \mu_0 - \mu_1 = 0$ with complementary slackness conditions $\mu_0\tilde{p}_0 = 0$, $\mu_1(\tilde{p}_0 - 1) = 0$ and $\mu_0, \mu_1 \geq 0$.

In the internal solution we have $\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=\tilde{p}_0^*} = 0$. The complementary slackness conditions imply that if $-\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=0} \geq 0$ then $\tilde{p}_0^* = 0$. If $\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=1} \geq 0$ then $\tilde{p}_0^* = 1$.

The first condition is equivalent to:

$$\begin{aligned} -\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=0} \geq 0 \iff & -p_0 - \frac{1}{\gamma}\Big(1 - P(s=H|p_0)\eta - P(s=L|p_0)\lambda\eta\Big)(u_H - u_L) \geq 0 \\ \iff & 1 - P(s=H|p_0)\eta - P(s=L|p_0)\lambda\eta \leq \frac{-p_0\gamma}{u_H - u_L} \end{aligned} \tag{17}$$

The second complementary slackness condition is equivalent to:

$$\begin{aligned} \frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=1} \geq 0 \iff & -1 + p_0 + \frac{1}{\gamma}\Big(1 - P(s=H|p_0)\eta - P(s=L|p_0)\lambda\eta\Big)(u_H - u_L) \geq 0 \\ \iff & 1 - P(s=H|p_0)\eta - P(s=L|p_0)\lambda\eta \geq \frac{(1-p_0)\gamma}{u_H - u_L} \end{aligned} \tag{18}$$

Let us denote $\frac{\gamma}{(u_H - u_L)}$ with q and set $\eta = 1$. The first complementary slackness condition can be rewritten as:

$$-p_0 - \frac{1}{q}\Big(1 - P(s = H|p_0) - P(s = L|p_0)\lambda\Big) \geq 0 \iff$$

$$\frac{1 - P(s = H|p_0) + qp_0}{P(s = L|p_0)} \leq \lambda \iff 1 + \frac{qp_0}{1 - P(s = H|p_0)} \leq \lambda$$

Substituting $P(s = H|p_0) = (2c - 1)p_0 + (1 - c)$, we get:

$$1 + \frac{qp_0}{(1 - 2c)p_0 + c} \leq \lambda \tag{19}$$

Rewriting the second complementary slackness condition in a similar way gives us:

$$1 - \frac{q(1 - p_0)}{(1 - 2c)p_0 + c} \geq \lambda \tag{20}$$

Thus, we obtain the conditions for the interior as well as the boundary solutions:

$$\begin{cases} \tilde{p}_0^* = 1 & \text{for} \quad \lambda \leq 1 - \dfrac{q(1 - p_0)}{(1 - 2c)p_0 + c} \\ \tilde{p}_0^* \in (0, 1) & \text{for} \quad 1 - \dfrac{q(1 - p_0)}{(1 - 2c)p_0 + c} \leq \lambda \leq 1 + \dfrac{qp_0}{(1 - 2c)p_0 + c} \\ \tilde{p}_0^* = 0 & \text{for} \quad \lambda \geq 1 + \dfrac{qp_0}{(1 - 2c)p_0 + c} \end{cases}$$

It is worth noting that, since $p_0 \in [0, 1]$ and $c \in (\frac{1}{2}, 1)$, the denominator in (19) and (20) is always positive. The LHS of (19) is always greater than 1, whereas the LHS of (20) is smaller than 1. In the table below, I show, for three different values of the unmanipulated belief, the lower and upper bound of the value of λ for which we have interior solution. When p_0 increases, the interval shifts to the right. Moreover, if we assume the manipulation costs to be large enough, the upper/lower bound will not be binding for any λ within reasonable range. The only agents that will hit the boundary are those with 1). low p_0 and high λ, and 2). high p_0 and low λ.

Table 10: Lower and upper bound for λ to ensure the interior solution given p_0.

	Interior Solution	
Unmanipulated Belief	Lower bound for λ	Upper bound for λ
$p_0 = 0$	$1 - \frac{q}{c}$	1
$p_0 = \frac{1}{2}$	$1 - q$	$1 + q$
$p_0 = 1$	1	$1 + \frac{q}{1-c}$

Unexpected Signals

Let μ_0 be the Kuhn-Tucker multiplier on the constraint $-\tilde{p}_0 \leq 0$ and μ_1 be the multiplier on the constraint $\tilde{p}_0 \leq 1$. The Lagrangian is given by

$$\mathcal{L} = U(\tilde{p}_0) + \mu_0(\tilde{p}_0) - \mu_1(\tilde{p}_0 - 1)$$

The optimality condition is $\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0} + \mu_0 - \mu_1 = 0$ with complementary slackness conditions $\mu_0\tilde{p}_0 = 0$, $\mu_1(\tilde{p}_0 - 1) = 0$ and $\mu_0, \mu_1 \geq 0$.

In the internal solution we have $\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0 = \tilde{p}_0^*} = 0$. The complementary slackness conditions imply that if $-\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=0} \geq 0$ then $\tilde{p}_0^* = 0$. If $\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=1} \geq 0$ then $\tilde{p}_0^* = 1$.

The first condition is equivalent to:

$$\begin{aligned} -\frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=0} \geq 0 \iff & -p_0 - \frac{1}{\gamma}\Big(1 - \epsilon P(s=H|p_0)\eta - \epsilon P(s=L|p_0)\lambda\eta\Big)(u_H - u_L) \geq 0 \\ \iff & 1 - \epsilon P(s=H|p_0)\eta - \epsilon P(s=L|p_0)\lambda\eta \leq \frac{-p_0\gamma}{u_H - u_L} \end{aligned} \tag{21}$$

The second complementary slackness condition is equivalent to:

$$\begin{aligned} \frac{\mathrm{d}\,U(\tilde{p}_0)}{\mathrm{d}\,\tilde{p}_0}|_{\tilde{p}_0=1} \geq 0 \iff & -1 + p_0 + \frac{1}{\gamma}\Big(1 - \epsilon P(s=H|p_0)\eta - \epsilon P(s=L|p_0)\lambda\eta\Big)(u_H - u_L) \geq 0 \\ \iff & 1 - \epsilon P(s=H|p_0)\eta - \epsilon P(s=L|p_0)\lambda\eta \geq \frac{(1-p_0)\gamma}{u_H - u_L} \end{aligned} \tag{22}$$

Let us denote $\frac{\gamma}{(u_H - u_L)}$ with q and set $\eta = 1$. The first complementary slackness condition can be rewritten as:

$$-p_0 - \frac{1}{q}\Big(1 - \epsilon P(s = H|p_0) - \epsilon P(s = L|p_0)\lambda\Big) \geq 0 \iff$$

$$\frac{1 - \epsilon P(s = H|p_0) + qp_0}{\epsilon P(s = L|p_0)} \leq \lambda \iff \frac{1 - \epsilon P(s = H|p_0) + \epsilon - \epsilon + qp_0}{\epsilon(1 - P(s = H|p_0))} \leq \lambda$$

$$1 + \frac{1 - \epsilon}{\epsilon P(s = L|p_0)} + \frac{qp_0}{\epsilon P(s = L|p_0)} \leq \lambda$$

Rewriting the second complementary slackness condition in a similar way gives us:

$$1 + \frac{1 - \epsilon}{\epsilon P(s = L|p_0)} - \frac{q(1 - p_0)}{\epsilon P(s = L|p_0)} \geq \lambda \tag{23}$$

The conditions for the interior and boundary solutions (denoting $\mathrm{P}(s = L|p_0)$ with P_L):

$$\begin{cases} \tilde{p}_0^* = 1 & \text{for} \quad \lambda \leq 1 + \dfrac{1 - \epsilon}{\epsilon P_L} - \dfrac{q(1 - p_0)}{\epsilon P_L} \\ \tilde{p}_0^* \in (0, 1) & \text{for} \quad 1 + \dfrac{1 - \epsilon}{\epsilon P_L} - \dfrac{q(1 - p_0)}{\epsilon P_L} \leq \lambda \leq 1 + \dfrac{1 - \epsilon}{\epsilon P_L} + \dfrac{qp_0}{\epsilon P_L} \\ \tilde{p}_0^* = 0 & \text{for} \quad \lambda \geq 1 + \dfrac{1 - \epsilon}{\epsilon P_L} + \dfrac{qp_0}{\epsilon P_L} \end{cases}$$

Substituting $P(s = L|p_0) = (1-2c)p_0 + c$, we can calculate the boundaries for different values of unmanipulated beliefs. In the table below, I gather the lower and upper bound of the value of λ for which we have interior solution in the cases $p_0 = 0$, $p_0 = \frac{1}{2}$, and $p_0 = 1$. It is important to note that the bounds depend on the values of ϵ. For ϵ close to 1, the upper and lower bound will be close to the case of an agent expecting a signal.

For ϵ closer to 0, however, the term $\frac{1-\epsilon}{\epsilon c}$ is very high—the upper bound becomes out of reach for any λ within a reasonable range. Moreover, for ϵ close to zero and a sufficiently high p_0, also the lower bound becomes very high. See, for example, the lower and upper bound for λ if $p_0 = 1$, which are presented in the last row in Table 11. In these cases, there is no λ large enough to ensure an interior solution. Both loss-averse and gain-seeking agents will end up choosing the corner solution $\tilde{p}_0^* = 1$.

Table 11: Lower and upper bound for λ to ensure the interior solution given p_0.

	Interior Solution	
Unmanipulated Belief	Lower bound for λ	Upper bound for λ
$p_0 = 0$	$1 + \frac{1-\epsilon}{\epsilon c} - \frac{q}{\epsilon c}$	$1 + \frac{1-\epsilon}{\epsilon c}$
$p_0 = \frac{1}{2}$	$1 + \frac{2(1-\epsilon)}{\epsilon c} - \frac{q}{\epsilon}$	$1 + \frac{2(1-\epsilon)}{\epsilon c} + \frac{q}{\epsilon}$
$p_0 = 1$	$1 + \frac{1-\epsilon}{\epsilon c}$	$1 + \frac{1-\epsilon}{\epsilon(1-c)} + \frac{q}{\epsilon(1-c)}$

In the extreme case when $\epsilon = 0$, the agent chooses $\tilde{p}_0$ to maximize the following:

$$U_0 = u(\tilde{p}_0) - \frac{\gamma}{2}(\tilde{p}_0 - p_0)^2, \tag{24}$$

subject to the following inequality constraints: $\tilde{p}_0 \leq 1$ and $\tilde{p}_0 \geq 0$. Neither the interior solution $\tilde{p}_0^* = p_0 + \frac{1}{\gamma}(u_H - u_L)$ nor the boundary solution $\tilde{p}_0^* = 1$ depend on λ.

B.2 Proposition 1

Let us denote $P(s = H|p_0)$ with P_H and $P(s = L|p_0)$ with P_L.
$\tilde{p_0}^* < \hat{p_0}^*$ is equivalent to:

$$p_0 + \frac{1}{\gamma}\Big(1 - P_H\eta - P_L\lambda\eta\Big)(u_H - u_L) < p_0 + \frac{1}{\gamma}\Big(1 - \epsilon P_H\eta - \epsilon P_L\lambda\eta\Big)(u_H - u_L) \iff$$

$$\Big(1 - P_H\eta - P_L\lambda\eta\Big) < \Big(1 - \epsilon P_H\eta - \epsilon P_L\lambda\eta\Big) \iff -P_H\eta - P_L\lambda\eta < -\epsilon P_H\eta - \epsilon P_L\lambda\eta \iff$$

$$(1 - \epsilon)\eta(P_H + P_L\lambda) > 0. \tag{25}$$

Since $\epsilon \in (0, 1)$, $P_H \in [0, 1]$, $P_L \in [0, 1]$, $\eta > 0$, $\lambda > 0$, the LHS is always strictly greater than zero.

B.3 Proposition 3

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 $(1-c)$. In this case, we can rewrite the probability of receiving a "good" signal as: $P(s = H|p_0) = (2c-1)p_0 + 1 - c$. The probability of receiving a "bad" signal is: $P(s = L|p_0) = c - (2c-1)p_0$. Then, we can rewrite the optimal belief as:

$$\tilde{p_0}^* \;=\; p_0 \;+\; \frac{1}{\gamma}\Bigg(1 - ((2c-1)p_0 + 1 - c)\;\eta - (c - (2c-1)p_0)\;\lambda\eta\Bigg)(u_H - u_L)$$

Taking the derivative with respect to λ gives us:

$$\frac{\partial \tilde{p_0}^*}{\partial \lambda} = \;\frac{1}{\gamma}\Bigg((c - (2c-1)p_0)\;\eta\Bigg)(u_H - u_L)$$

The partial derivative with respect to p_0:

$$\frac{\partial \partial \tilde{p_0}^*}{\partial \lambda \partial p_0} = \;\frac{1}{\gamma}(2c-1)\;\eta(u_H - u_L) \tag{26}$$

Since $c > 0.5$, $\gamma > 0$, $\eta > 0$, and $u_H > u_L$, (26) is strictly greater than zero.

B.4 Over- and Under-confidence in Control

Low-ability Agents

Assuming that $\tilde{p}_0^*$ is the interior solution, the unmanipulated belief is degenerate $p_0 = 0$ and perfectly captures ability, the agent's bias is positive if and only if:

$$\begin{aligned} &\hat{p_0}^* \;-\; p_0 \;> 0 \iff \frac{1}{\gamma}\Bigg(1 - \epsilon P_H\eta - \epsilon P_L\lambda\eta\Bigg)(u_H - u_L) > 0 \iff \\ &1 - \epsilon P_H\eta - \epsilon P_L\lambda\eta > 0 \iff \frac{1 - \epsilon P_H\eta}{\epsilon P_L\eta} > \lambda \end{aligned} \tag{27}$$

Since the agent is of low ability ($p_0 = 0$), he receives a signal L with probability $P_L = c$, and a signal H with probability $P_H = 1-c$, where $c > \frac{1}{2}$. For $\eta = 1$, the above inequality reduces to: $1 + \frac{1-\epsilon}{\epsilon c} > \lambda$.

High-ability Agents

Assuming that the unmanipulated belief is degenerate $p_0 = 1$, the agent's bias is negative (indicating underconfidence) if and only if:

$$\begin{aligned} &\hat{p_0}^* \;-\; p_0 \;< 0 \iff \frac{1}{\gamma}\Big(1 - \epsilon P_H \eta - \epsilon P_L \lambda \eta\Big)(u_H - u_L) < 0 \iff \\ &1 - \epsilon P_H \eta - \epsilon P_L \lambda \eta < 0 \iff \frac{1 - \epsilon P_H}{\epsilon P_L} < \lambda \end{aligned} \tag{28}$$

Since the agent is of high ability ($p_0 = 1$), he receives a signal H with probability $P_H = c$, and a signal L with probability $P_L = 1 - c$, where $c > \frac{1}{2}$. For $\eta = 1$, we have:

$$1 + \frac{1 - \epsilon}{\epsilon(1 - c)} < \lambda \tag{29}$$

Thus, Proposition B.1 in the Control condition takes the following form:

Proposition B.1 (Control)

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 of low-ability and $\lambda < 1 + \frac{1-\epsilon}{\epsilon c}$,
ii) an agent is underconfident if and only if he is of high-ability and $\lambda > 1 + \frac{1-\epsilon}{\epsilon(1-c)}$.
iii) an agent is unbiased if and only if he is low-ability and $\lambda \geq 1 + \frac{1-\epsilon}{\epsilon c}$, or he is high-ability and $\lambda \leq 1 + \frac{1-\epsilon}{\epsilon c}$.

For ϵ close to one, the conditions become close to the case of expecting a signal. For ϵ close to zero, however, the condition in Proposition B.1 ii) becomes to restrictive to hold for any reasonable value of λ. At the same time, the condition in Proposition B.1 i) is weakened and is likely to be fulfilled for any λ. In consequence, for ϵ close to zero, all low-ability agents should be overconfident, whereas no high-ability agent–underconfident.

Second, the agent bias takes the following form:

$$\hat{p_0}^* \;-\; p_0 \;=\; \frac{1}{\gamma}\Big(1-\epsilon P_H\eta-\epsilon P_L\lambda\eta\Big)(u_H-u_L) \;=$$

$$=\; \frac{1}{\gamma}\Big(1-\epsilon\big((2c-1)p_0+1-c\big)\eta-\epsilon\big(c-(2c-1)p_0\big)\lambda\eta\Big)(u_H-u_L) \;=\quad \text{*assuming } \eta=1$$

$$=\; \frac{1}{\gamma}\Big(1-\epsilon-\epsilon(1-\lambda)(2c-1)p_0+\epsilon c(1-\lambda)\Big)(u_H-u_L) \;=$$

$$=\; \frac{1}{\gamma}\Big(1-\epsilon+\epsilon(1-\lambda)\underbrace{\big(c-(2c-1)p_0\big)}_{P_L}\Big)(u_H-u_L). \tag{30}$$

The bias depends on ϵ, λ and P_L. The latter further reduces to c for low-ability individuals ($p_0 = 0$) and to $(1-c)$ for high-ability subjects (when $p_0 = 1$). Proposition B.2 takes the same form in the Control as in the Treatment condition:

Proposition B.2 (Control)

i) The absolute bias is larger for low-ability than for high-ability agents.

ii) The absolute bias is decreasing in the loss aversion parameter λ for low-ability agents, and increasing in λ for high-ability individuals.

iii) λ has a stronger impact on the absolute bias for low-ability agents.

However, both the difference in the absolute bias between high- and low-ability agents and the effects of λ are attenuated due to $\epsilon \in (0,1)$. For ϵ close to 1, the effects will be similar to the Treatment condition. For a lower ϵ, however, the attenuation will be more pronounced, up to the point where there is no difference in bias, which becomes independent of λ.

C Robustness Checks

C.1 Restricted sample (high ϵ).

In this section, I present the results based on different restricted samples. In the first sample (Restricted sample 1.a), I exclude all high-ability non-loss-averse and low-ability loss-averse subjects from the two conditions.

The exclusion criteria used to construct the second sample (Restricted sample 2.a) are much weaker. Instead of excluding all loss-averse participants with the lowest rank, I exclude only participants with the highest measure of loss aversion (values above 7, on a scale from 1 to 9). Similarly, instead of excluding all non-loss-averse subjects with the highest rank, I exclude those with the lowest measure of loss aversion (values below 3). I consider subjects ranked 9 and 10 to have the lowest rank, and those ranked 1 and 2 to have the highest rank. The same exclusion criteria are used in the Treatment and the Control condition. All in all, I excluded 9 participants (4% of the sample), a number much smaller than 30 participants excluded in the first sample. The results are presented in Table 12, 13, and 14.

Table 12: The effect of treatment on mean beliefs.

	Restricted sample 1.a		Restricted sample 2.a	
Treatment	-0.249	-0.189	-0.149	-0.035
	(0.233)	(0.228)	(0.230)	(0.224)
Ability		0.161***		0.175***
		(0.049)		(0.042)
Const.	6.506***	5.604***	6.415***	5.419***
	(0.163)	(0.316)	(0.161)	(0.286)
N	197	197	218	218

The dependent variable is the mean belief revealed in Belief Elicitation I.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: The effect of treatment and loss aversion on mean beliefs.

	Restricted sample 1.a		Restricted sample 2.a	
	(1)	(2)	(1)	(2)
Treatment	1.240*	1.114*	1.428**	1.354**
	(0.664)	(0.648)	(0.687)	(0.662)
Loss Aversion	0.075	0.020	0.039	0.002
	(0.089)	(0.088)	(0.091)	(0.088)
Treatment × Loss Aversion	-0.311**	-0.271**	-0.323**	-0.283**
	(0.131)	(0.128)	(0.134)	(0.130)
Ability		0.165***		0.175***
		(0.049)		(0.042)
N	197	197	218	218

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

Table 14: The effect of treatment, loss aversion, and unmanipulated beliefs.

	Restricted sample 1.a		Restricted sample 2.a	
	(1)	(2)	(1)	(2)
Treatment	-0.111	0.461	0.041	0.658
	(0.227)	(0.773)	(0.220)	(0.766)
Loss Aversion	-0.477***	-0.325*	-0.528***	-0.367**
	(0.173)	(0.185)	(0.157)	(0.171)
Ability	-0.126	-0.188	-0.173	-0.212
	(0.140)	(0.147)	(0.134)	(0.137)
Ability × LA	0.065**	0.063**	0.073***	0.068**
	(0.028)	(0.028)	(0.026)	(0.026)
Treatment × LA		-0.279**		-0.270**
		(0.128)		(0.128)
Treatment × Ability		0.139		0.127
		(0.098)		(0.082)
N	197	197	218	218

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

C.2 Restricted sample (low ϵ).

In this section, I present the results based on two different restricted samples. The exclusion criteria differ in the Treatment and the Control condition. In the Control, I define them assuming ϵ close to zero, see Appendix B.1.

In the first sample (Restricted sample 1.b), I exclude all high-ability participants (whose rank was Rank 1) in the Control condition, and all high-ability non-loss-averse and low-ability loss-averse subjects in the Treatment condition (in this condition, the exclusion criteria are the same as in the Restricted sample 1.a).

The exclusion criteria in the second sample (Restricted sample 2.b) are weakened in the Treatment condition in the same way as in the previous section: I exclude only low-ability participants with the highest measure of loss aversion (values above 7, on a scale from 1 to 9) and high-ability with the lowest measure of loss aversion (values below 3). The exclusion criteria in Control remain the same: I exclude all high-ability participants (whose rank was Rank 1). The results are gathered in Table 15, 16, and 17.

Table 15: The effect of treatment on mean beliefs.

	Restricted sample 1.b		Restricted sample 2.b	
Treatment	-0.113	-0.141	-0.103	-0.096
	(0.236)	(0.228)	(0.234)	(0.223)
Ability		0.191***		0.210***
		(0.049)		(0.045)
Const.	6.369***	5.401***	6.369***	5.306***
	(0.164)	(0.293)	(0.167)	(0.277)
N	199	199	210	210

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

Table 16: The effect of treatment and loss aversion on mean beliefs.

	Restricted sample 1.b		Restricted sample 2.b	
	(1)	(2)	(1)	(2)
Treatment	0.868	0.940	1.137*	1.145*
	(0.675)	(0.649)	(0.686)	(0.653)
Loss Aversion	-0.032	-0.029	-0.032	-0.029
	(0.090)	(0.086)	(0.091)	(0.086)
Treatment × Loss Aversion	-0.203	-0.224*	-0.252*	-0.252**
	(0.132)	(0.127)	(0.134)	(0.127)
Ability		0.197***		0.209***
		(0.048)		(0.044)
N	199	199	210	210

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

Table 17: The effect of treatment, loss aversion, and unmanipulated beliefs.

	Restricted sample 1.b		Restricted sample 2.b	
	(1)	(2)	(1)	(2)
Treatment	-0.155	0.661	-0.073	0.865
	(0.223)	(0.793)	(0.218)	(0.783)
Loss Aversion	-0.495***	-0.423***	-0.486***	-0.380**
	(0.153)	(0.158)	(0.149)	(0.155)
Ability	-0.139	-0.222	-0.119	-0.179
	(0.137)	(0.146)	(0.137)	(0.143)
Ability × LA	0.071**	0.080***	0.067**	0.071***
	(0.027)	(0.027)	(0.027)	(0.027)
Treatment × LA		-0.275**		-0.275**
		(0.126)		(0.126)
Treatment × Ability		0.096		0.077
		(0.096)		(0.087)
N	199	199	210	210

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

C.3 Restricted sample (high or low ϵ).

In this section, I present the results based on two different restricted samples. The exclusion criteria differ in the Treatment and the Control condition. In the Control, the exclusion criteria are a combination of those used in the previous subsections.

In the first sample (Restricted sample 1.c), I exclude all high-ability participants (whose rank was Rank 1) in the Control condition, as well as all high-ability non-loss-averse and low-ability loss-averse subjects in the Treatment and the Control condition.

The exclusion criteria in the second sample (Restricted sample 2.c) are weakened in the Treatment condition in the same way as in the previous section: I exclude only low-ability participants with the highest measure of loss aversion (values above 7, on a scale from 1 to 9) and high-ability with the lowest measure of loss aversion (values below 3). In the Control condition, I exclude those who fulfill these criteria as well as all high-ability participants (whose rank was Rank 1). The results are gathered in Table 18, 19, and 20.

Table 18: The effect of treatment on mean beliefs.

	Restricted sample 1.c		Restricted sample 2.c	
Treatment	-0.225	-0.200	-0.118	-0.103
	(0.237)	(0.231)	(0.236)	(0.226)
Ability		0.167***		0.203***
		(0.052)		(0.046)
Const.	6.481***	5.586***	6.384***	5.349***
	(0.167)	(0.321)	(0.171)	(0.287)
N	192	192	205	205

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

Table 19: The effect of treatment and loss aversion on mean beliefs.

	Restricted sample 1.c		Restricted sample 2.c	
	(1)	(2)	(1)	(2)
Treatment	1.157*	1.102*	1.385**	1.236*
	(0.673)	(0.655)	(0.697)	(0.668)
Loss Aversion	0.054	0.020	0.023	-0.009
	(0.094)	(0.092)	(0.096)	(0.092)
Treatment × Loss Aversion	-0.290**	-0.271**	-0.307**	-0.273**
	(0.134)	(0.131)	(0.137)	(0.131)
Ability		0.173***		0.202***
		(0.051)		(0.045)
N	192	192	205	205

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

Table 20: The effect of treatment, loss aversion, and unmanipulated beliefs.

	Restricted sample 1.c		Restricted sample 2.c	
	(1)	(2)	(1)	(2)
Treatment	-0.140	0.640	-0.029	0.919
	(0.228)	(0.799)	(0.222)	(0.782)
Loss Aversion	-0.520***	-0.366*	-0.554***	-0.411**
	(0.183)	(0.193)	(0.165)	(0.174)
Ability	-0.150	-0.201	-0.176	-0.235
	(0.147)	(0.151)	(0.148)	(0.153)
Ability × LA	0.074**	0.073**	0.080***	0.081***
	(0.031)	(0.031)	(0.029)	(0.029)
Treatment × LA		-0.290**		-0.292**
		(0.130)		(0.129)
Treatment × Ability		0.110		0.089
		(0.101)		(0.090)
N	192	192	205	205

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

C.4 Restricted sample: excluding all with the highest/lowest ability

In this section, I present the results from Table 2, 3, and 4 based on a different restricted sample. I exclude participants with the highest and the lowest rank (Rank 1 and 10), regardless of their gain-loss attitudes.

Figure 9: The mean belief depending on loss aversion in the two conditions.

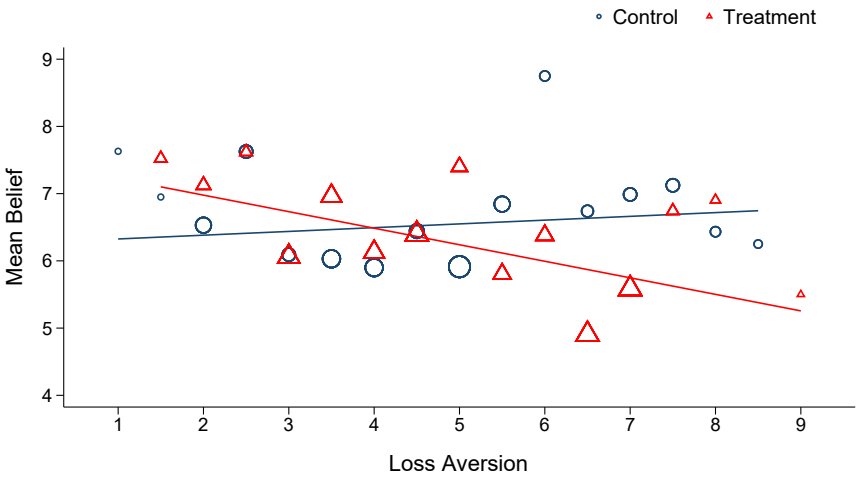


Table 21: The effect of treatment on mean beliefs.

	All subjects		Restricted sample	
Treatment	-0.141	-0.026	-0.263	-0.169
	(0.225)	(0.218)	(0.248)	(0.243)
Ability		0.178***		0.217***
		(0.040)		(0.066)
Const.	6.408***	5.402***	6.533***	5.243***
	(0.157)	(0.272)	(0.177)	(0.429)
N	227	227	173	173

Table 22: The effect of treatment and loss aversion on mean beliefs.

	All subjects		Restricted sample (R)	
	(1)	(2)	(1)	(2)
Treatment	1.038	1.049*	1.200*	1.253**
	(0.657)	(0.632)	(0.696)	(0.676)
Loss Aversion	-0.021	-0.022	0.056	0.048
	(0.085)	(0.082)	(0.096)	(0.093)
Treatment × Loss Aversion	-0.238*	-0.218*	-0.302**	-0.294**
	(0.127)	(0.123)	(0.136)	(0.132)
Ability		0.173***		0.216***
		(0.040)		(0.065)
N	227	227	173	173

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

Table 23: The effect of treatment, loss aversion, and unmanipulated beliefs.

	All subjects		Restricted sample	
	(1)	(2)	(1)	(2)
Treatment	0.011	0.507	-0.141	0.346
	(0.215)	(0.767)	(0.240)	(0.992)
Loss Aversion	-0.403***	-0.297**	-0.494***	-0.344
	(0.136)	(0.150)	(0.203)	(0.212)
Ability	-0.083	-0.107	-0.110	-0.196
	(0.118)	(0.123)	(0.173)	(0.187)
Ability × LA	0.053**	0.050**	0.070***	0.070***
	(0.023)	(0.023)	(0.034)	(0.034)
Treatment × LA		-0.194		-0.290**
		(0.122)		(0.131)
Treatment × Ability		0.083		0.157
		(0.079)		(0.130)
N	227	227	173	173

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

C.5 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 10, 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, I 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 24, 25, and 26 correspond to Tables 2, 3, and 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).

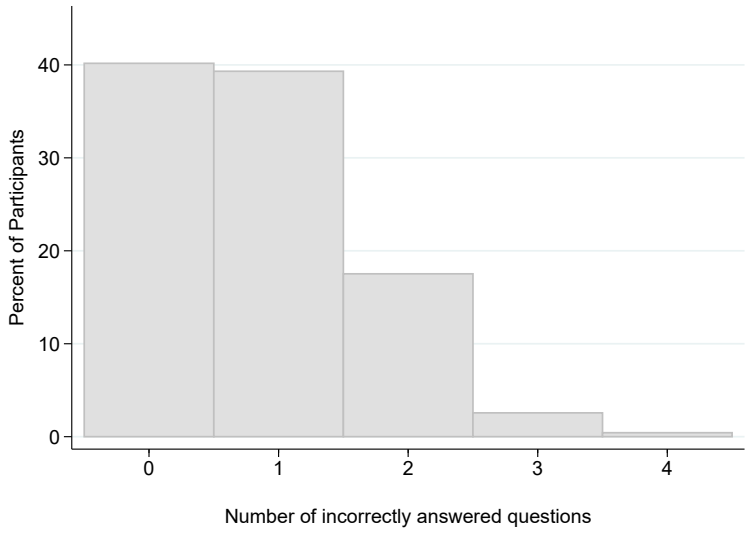


Figure 10: Control Questions

Table 24: 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 **??**). "R": the remaining subjects (all subjects minus boundary cases). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 25: 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 26: The effect of treatment, loss aversion, and unmanipulated beliefs.

	All subjects		Restricted sample (R)	
	(1)	(2)	(1)	(2)
Treatment	0.014	0.347	-0.111	0.295
	(0.214)	(0.771)	(0.227)	(0.780)
Loss Aversion	-0.405***	-0.317**	-0.507***	-0.379**
	(0.137)	(0.150)	(0.175)	(0.187)
Ability	-0.065	-0.101	-0.129	-0.195
	(0.118)	(0.123)	(0.142)	(0.150)
Ability × Loss Aversion	0.050**	0.048**	0.067**	0.066**
	(0.023)	(0.023)	(0.029)	(0.029)
Treatment × Loss Aversion		-0.168		-0.244*
		(0.122)		(0.130)
Treatment × Ability		0.091		0.142
		(0.079)		(0.098)
N	234	234	203	203

"Loss Aversion" is a dummy based on loss aversion measure (1 if average response to the hypothetical questions was above 5). "Ability" denotes position in the IQ test distribution (taking values from 1 to 10, with 10 denoting subjects with the highest scores). All specifications include a constant (omitted for clarity). "All" refers to the sample of all subjects and "R" to the restricted sample.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C.6 Excluding the most biased subjects

One might wonder whether the results might be driven by a small number of people with extremely inaccurate beliefs about themselves. In Tables 27, 28, and 29, I show how the results from Tables 2, 3, and 4 change if I exclude participants with the highest absolute bias. I exclude subjects 1) who were among 5% subjects with the highest bias or 5% subjects with the lowest 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 similar to the ones obtained without the restrictions.

Table 27: 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 28: 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 27.

Table 29: The effect of treatment, loss aversion, and unmanipulated beliefs.

	Restricted sample I		Restricted sample II	
	(1)	(2)	(1)	(2)
Treatment	0.082	0.940	-0.111	0.282
	(0.198)	(0.727)	(0.208)	(0.766)
Loss Aversion	-0.356***	-0.222	-0.331**	-0.233
	(0.131)	(0.143)	(0.135)	(0.147)
Ability	0.065	0.057	0.007	-0.021
	(0.114)	(0.118)	(0.117)	(0.121)
Ability × Loss Aversion	0.044**	0.039*	0.045**	0.042*
	(0.022)	(0.022)	(0.022)	(0.022)
Treatment × Loss Aversion		-0.241**		-0.179
		(0.111)		(0.118)
Treatment × Ability		0.059		0.089
		(0.076)		(0.079)
N	207	207	220	220

"Loss Aversion" takes values from 1 to 9 and denotes the average response to the hypothetical questions."Ability" denotes position in the test score distribution. "RS I" denotes the sample without 5% of subjects with the highest and 5% with the lowest bias (thus, I exclude subjects with extreme positive or negative bias). "RS II" includes 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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C.7 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 30, 31, and 32, 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 33, 34, and 35, 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 30: 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 **??**). "R": the remaining subjects (all subjects minus boundary cases). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 31: 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 32: The effect of treatment, loss aversion, and unmanipulated beliefs.

	All subjects		Restricted sample (R)	
	(1)	(2)	(1)	(2)
Treatment	0.039	0.315	-0.072	0.268
	(0.219)	(0.782)	(0.232)	(0.794)
Loss Aversion	-0.445***	-0.351**	-0.528***	-0.403**
	(0.139)	(0.153)	(0.177)	(0.190)
Ability	-0.101	-0.135	-0.143	-0.207
	(0.121)	(0.125)	(0.143)	(0.151)
Ability × Loss Aversion	0.057**	0.054**	0.070**	0.069**
	(0.023)	(0.023)	(0.029)	(0.029)
Treatment × Loss Aversion		-0.169		-0.230*
		(0.124)		(0.132)
Treatment × Ability		0.102		0.140
		(0.081)		(0.100)
N	227	227	197	197

"Loss Aversion" is a dummy based on loss aversion measure (1 if average response to the hypothetical questions was above 5). "Ability" denotes position in the IQ test distribution (taking values from 1 to 10, with 10 denoting subjects with the highest scores). All specifications include a constant (omitted for clarity). "All" refers to the sample of all subjects and "R" to the restricted sample.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 33: The effect of treatment on the 1^{st} and 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 34: The effect of treatment and loss aversion on the 1^{st} and 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 35: The effect of treatment, loss aversion, and unmanipulated beliefs.

Dependent variable:	the 1^{st} quartile		the 3^{rd} quartile	
	(1)	(2)	(1)	(2)
Treatment	-0.083	0.520	-0.193	0.525
	(0.224)	(0.765)	(0.243)	(0.830)
Loss Aversion	-0.523***	-0.371**	-0.424**	-0.260
	(0.171)	(0.183)	(0.185)	(0.199)
Ability	-0.161	-0.220	-0.103	-0.159
	(0.138)	(0.145)	(0.150)	(0.158)
Ability × Loss Aversion	0.075***	0.072**	0.060**	0.058*
	(0.028)	(0.028)	(0.030)	(0.030)
Treatment × Loss Aversion		-0.278**		-0.299**
		(0.127)		(0.138)
Treatment × Ability		0.133		0.130
		(0.097)		(0.105)
N	197	197	197	197

"Loss Aversion" is a dummy based on loss aversion measure (1 if average response to the hypothetical questions was above 5). "Ability" denotes position in the IQ test distribution (taking values from 1 to 10, with 10 denoting subjects with the highest scores). All specifications include a constant (omitted for clarity). "All" refers to the sample of all subjects and "R" to the restricted sample.
Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

D 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:

$$\begin{aligned} U_0 \;=\; u(\tilde{p}_0) \;+\; P(s=H|p_0) \underbrace{\left[\; u(p_1^H) \;+\; \eta \left(u(p_1^H) - u(\tilde{p}_0) \right) \right]}_{U_g \text{ (utility after a ``good'' signal } s=H)} + \\ + \;\; P(s=L|p_0) \underbrace{\left[\; u(p_1^L) \;+\; \eta \;\lambda \left( u(p_1^L) - u(\tilde{p}_0) \right) \right]}_{U_l \text{ (utility after a ``bad'' signal } s=L)} - \frac{\gamma}{2}(\tilde{p}_0 - p_0)^2, \end{aligned} \tag{31}$$

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.[27] 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 11) and were incentivized to report truthfully with binarized scoring rule (Hossain and Okui, 2013). While participants 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

[27] 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: 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).

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, I 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 F, I describe how to use subjects' responses to identify λ_i.

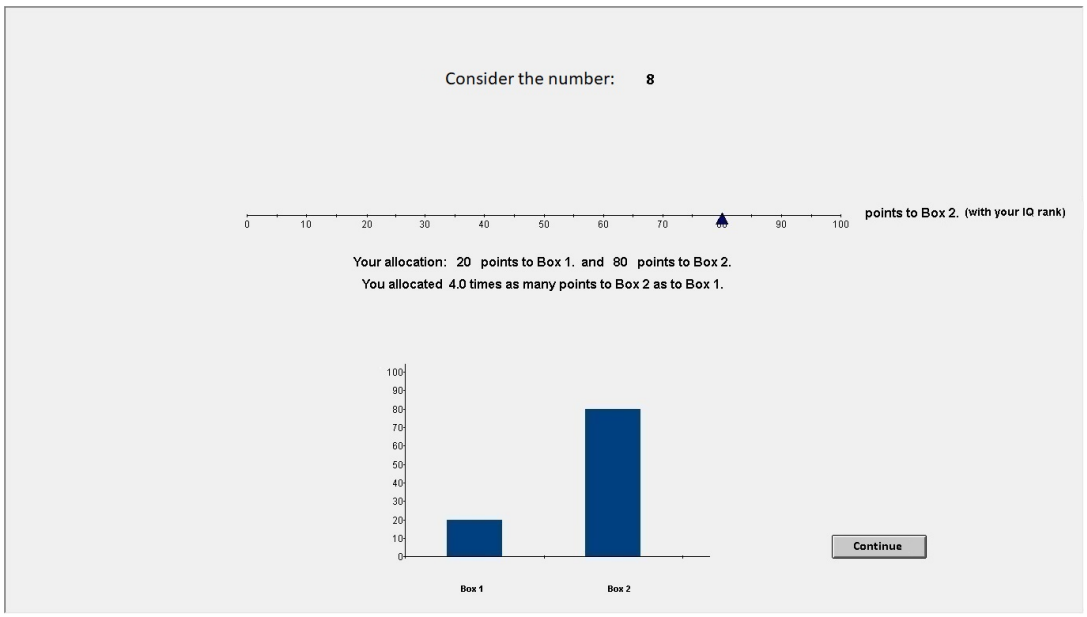


Figure 11: The screen-shot of the interface used in hypothetical questions.

D.1 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.[28] Participants in the 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.[29] 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.[30]

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

[29] 65% of subjects followed the profit-maximizing strategy in *every* decision they made.

[30] 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

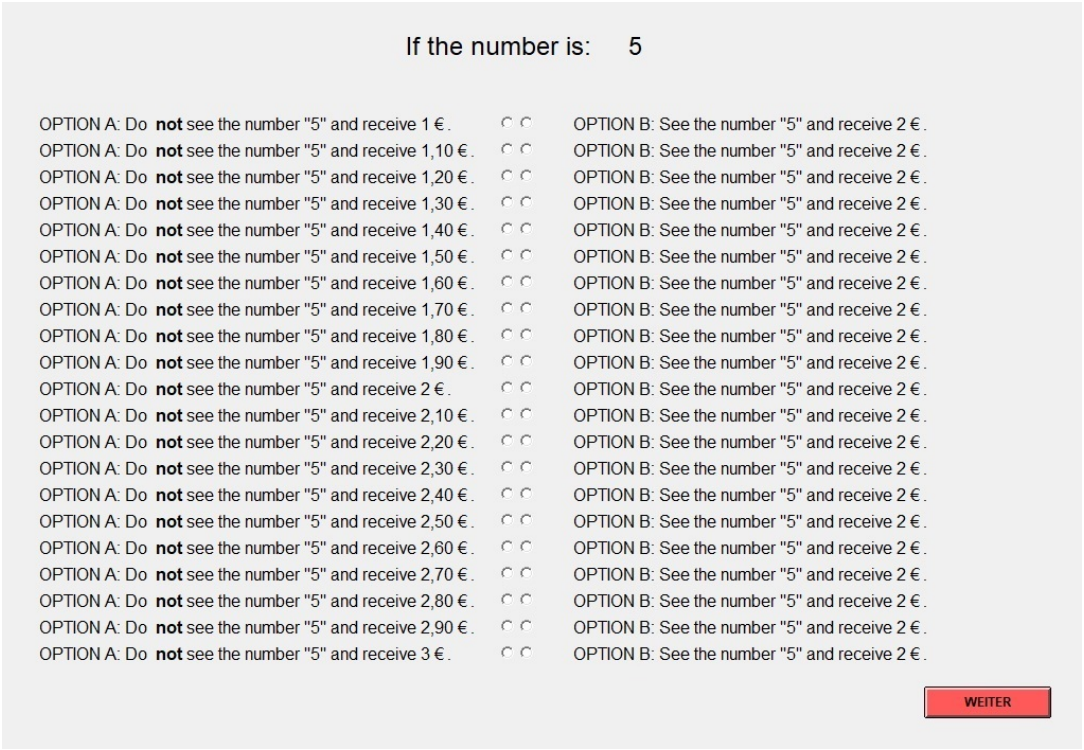
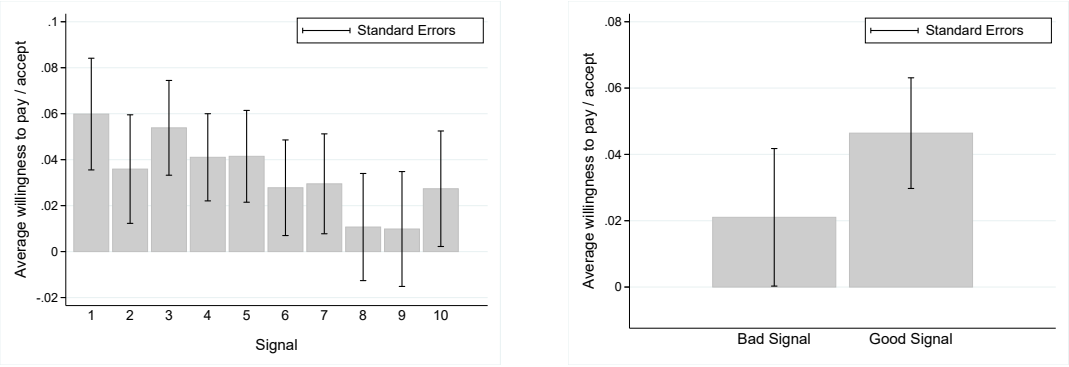


Figure 12: The screen-shot of the interface used by subjects in WTP.

Figure 13: Average willingness to pay for a signal.



(a) Signals 1-10

(b) "Good" vs "Bad" Signals

In Figure 13, 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, see Figure 12). 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 Option A was chosen.[31] 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.

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.

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

In Panel (a) in Figure 13, 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".[32] The difference is statistically significant and consistent with the theory – high signals are more desirable as they entail higher belief-based utility. In Panel (b) in Figure 13, 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 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 E.

D.1.1 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

[32] 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.

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.[33]

I define a person as "loss-averse" if she avoided at least two out of the three worst signals.[34] There were 53 participants of this type; they constitute 30% of the restricted sample of 178 participants.[35] The measure of loss aversion described above coincides with the loss aversion based on the hypothetical questions for 60% of subjects (a low percentage it is to be expected given the quality of the WTP data).

I conduct the same analysis as in Section 5.2. The most important results are gathered in Table 36 (complete analysis can be found in Appendix E). 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 using the second measure of loss aversion. The coefficient at the interaction term is negative, as predicted by the model, 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.[36] I also use Specification (1) from Table 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.

[33] In Appendix E, 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.

[34] 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 participants (23% of the restricted sample). The results are almost the same as presented in this section.

[35] Note that the 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 E. Moreover, as in the previous section, I exclude participants for whom the predictions of the model do not hold.

[36] 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 36: 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)*

E 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 E.1-E.4 in tables that correspond to Tables 2, 3, and 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.[37] 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 the 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 E.1, I show the results summarized in the previous section (Table 36 in Section C.1.1). 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 E.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 E.3, I 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

[37] 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.

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 E.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 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 and passes the threshold for significance only in Section E.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 $\times$ Loss Aversion" is negative and significant, as predicted by the theory.

E.1 Results based on coarse classification (three worst signals)

Table 37: 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 38: The effect of treatment and loss aversion on mean beliefs.

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 39: The effect of treatment, loss aversion, and unmanipulated beliefs.

	RS1		RS3	
	(1)	(2)	(1)	(2)
Treatment	-0.211	-0.690	-0.074	-0.148
	(0.226)	(0.559)	(0.242)	(0.630)
Loss Aversion	-0.802	-0.402	-0.922	-0.389
	(0.671)	(0.751)	(0.701)	(0.784)
Ability	0.145**	0.086	0.128**	0.106
	(0.056)	(0.076)	(0.064)	(0.086)
Ability × Loss Aversion	0.120	0.109	0.140	0.121
	(0.105)	(0.106)	(0.110)	(0.111)
Treatment × Loss Aversion		-0.688		-0.836
		(0.518)		(0.534)
Treatment × Ability		0.122		0.059
		(0.096)		(0.105)
N	202	202	178	178

"R1" denotes the sample without the boundary cases – those who do not fulfill 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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E.2 Results based on coarse classification (five worst signals)

Table 40: 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 41: The effect of treatment and loss aversion on mean beliefs.

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 42: The effect of treatment, loss aversion, and unmanipulated beliefs.

	RS1		RS3	
	(1)	(2)	(1)	(2)
Treatment	-0.232	-0.759	-0.092	-0.225
	(0.224)	(0.550)	(0.248)	(0.652)
Loss Aversion	-0.846	-0.369	-0.984	-0.414
	(0.669)	(0.741)	(0.731)	(0.800)
Ability	0.150***	0.087	0.136**	0.107
	(0.056)	(0.074)	(0.067)	(0.087)
Ability × LA	0.140	0.125	0.159	0.146
	(0.105)	(0.105)	(0.114)	(0.114)
Treatment × LA		-0.779		-0.964*
		(0.521)		(0.550)
Treatment × Ability		0.134		0.074
		(0.094)		(0.108)
N	202	202	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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E.3 Results based on coarse classification (five worst signals, average)

Table 43: 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 44: The effect of treatment and loss aversion on mean beliefs.

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 45: The effect of treatment, loss aversion, and unmanipulated beliefs.

	RS1		RS3	
	(1)	(2)	(1)	(2)
Treatment	-0.221	-0.827	-0.153	-0.442
	(0.223)	(0.547)	(0.251)	(0.654)
Loss Aversion	-0.829	-0.456	-0.984	-0.553
	(0.674)	(0.740)	(0.726)	(0.797)
Ability	0.153***	0.083	0.121*	0.079
	(0.056)	(0.074)	(0.068)	(0.089)
Ability × LA	0.126	0.118	0.158	0.148
	(0.104)	(0.104)	(0.113)	(0.113)
Treatment × LA		-0.687		-0.752
		(0.515)		(0.548)
Treatment × Ability		0.145		0.095
		(0.094)		(0.109)
N	203	203	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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E.4 Results based on the relative measure

Table 46: 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 47: The effect of treatment and loss aversion on mean beliefs.

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 48: The effect of treatment, loss aversion, and unmanipulated beliefs.

	RS1		RS3	
	(1)	(2)	(1)	(2)
Treatment	-0.226	-0.801	-0.139	-0.349
	(0.224)	(0.551)	(0.254)	(0.663)
Loss Aversion	-0.800	-0.407	-0.932	-0.455
	(0.686)	(0.766)	(0.739)	(0.824)
Ability	0.153***	0.086	0.126*	0.091
	(0.055)	(0.074)	(0.068)	(0.090)
Ability × LA	0.117	0.108	0.146	0.132
	(0.106)	(0.107)	(0.115)	(0.116)
Treatment × LA		-0.694		-0.780
		(0.529)		(0.562)
Treatment × Ability		0.138		0.079
		(0.095)		(0.111)
N	202	202	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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

F 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.[38] I argue that this assumption is not very demanding considering the 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:

$$p^x_1 = p^x_{B2} \cdot e_{i=x} + (1 - p^x_{B2}) \cdot p_0 \tag{32}$$

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

[38] 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 (31) 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.

stated belief $\tilde{p}_0$. Although I have no direct proof that p_0 and p_1^x are the "unmanipulated" belief, there is some evidence that people 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 (31).

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 my 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:

$$p_1^H = \frac{1}{|S|} \sum_{x \in S} \sum_{s \leq \bar{m}} p_{1,s}^x \tag{33}$$

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:

$$\delta_H = \frac{1}{|S|} \sum_{x \in S} \delta_x \tag{34}$$

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 (31), we have:

$$\begin{aligned}
\delta_H &= p_1^H u_H + (1 - p_1^H)u_L + \eta \left[p_1^H u_H + (1 - p_1^H)u_L - \Big(p_0 u_H + (1 - p_0)u_L \Big) \right], \\
\delta_L &= p_1^L u_H + (1 - p_1^L)u_L + \eta\lambda \left[p_1^L u_H + (1 - p_1^L)u_L - \Big(p_0 u_H + (1 - p_0)u_L \Big) \right].
\end{aligned} \tag{35}$$

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 (35), 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:

$$\lambda = \frac{(\delta_L - 1)(2p_1^H - p_0)}{(\delta_H - 1)(p_1^L - p_0)} - \frac{p_1^L}{p_1^L - p_0}, \tag{36}$$

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 (35) – 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.[39] If this was the case, the equations for δ_H and δ_L would take the form:

$$\begin{aligned}\delta_H &= \eta\left[p_1^H u_H + (1-p_1^H)u_L - \Big(p_0 u_H + (1-p_0)u_L\Big)\right],\\ \delta_L &= \eta\lambda\left[p_1^L u_H + (1-p_1^L)u_L - \Big(p_0 u_H + (1-p_0)u_L\Big)\right].\end{aligned} \tag{37}$$

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:

$$\lambda \;=\; \frac{\delta_L}{\delta_H}\;\frac{p_1^H - p_0}{p_1^L - p_0}, \tag{38}$$

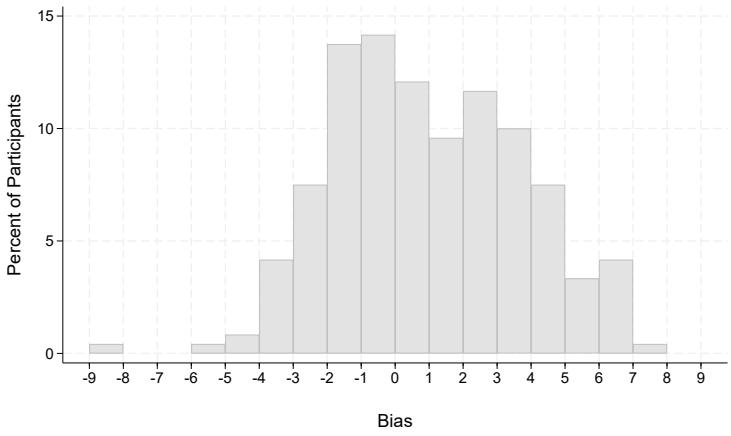
which is more straightforward than (36). Intuitively, the measure of loss aversion depends on the amount of money a subject is willing to forgo to avoid "bad" news (relative to the amount of money he would pay to avoid "good" news) weighted by how much his beliefs would move after each signal.

[39] 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.

G 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 5.1). The remaining subjects are classified as underconfident. There were 135 overconfident and 92 underconfident subjects in the sample.[40] The distribution of subjects' bias is presented in Figure 14. 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 14: Distribution of subjects' bias (based on Belief Elicitation I).



[40] 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.