Managing Self-Confidence: Theory and Experimental Evidence*

Markus M. Möbius
Microsoft Research, University of Michigan and NBER

Muriel Niederle
Stanford University and NBER

Paul Niehaus
UC San Diego and NBER

Tanya S. Rosenblat
University of Michigan

October 18, 2021

Abstract

We use a series of experiments to understand whether and how people's beliefs about their own abilities are biased relative to the Bayesian benchmark, and how these beliefs then affect behavior. We find that subjects systematically and substantially over-weight positive feedback relative to negative (asymmetry) and also update too little overall (conservatism). These biases are substantially less pronounced in an ego-free control experiment. Updating does retain enough of the structure of Bayes' rule to let us model it coherently in an optimizing framework in which, interestingly, asymmetry and conservatism emerge as complementary biases. We also find that exogenous changes in beliefs affect subjects' decisions to enter into a competition, and do so similarly for more and less biased subjects, suggesting that people cannot "undo" their biases when the time comes to decide.

JEL Classification: C91, C93, D83
Keywords: asymmetric belief updating, conservatism, information aversion

*We are grateful to the Department Editor and Associate Editor as well as to Nageeb Ali, Roland Benabou, Gary Chamberlain, Rachel Croson, Gordon Dahl, David Eil, Glenn Ellison, Asen Ivanov, John List, Justin Rao, Al Roth, Andrei Shleifer, Joel Sobel, Lise Vesterlund, Roberto Weber, and participants at numerous seminars for their feedback. Aislinn Bohren and Hanzhe Zhang provided outstanding research assistance. Jenő Pál provided very helpful comments on the final draft. Niederle and Rosenblat are grateful for the hospitality of the Institute for Advanced Study where part of this paper was written. We thank the National Science Foundation, Harvard University and Wesleyan University for financial support. Niehaus acknowledges financial support from an NSF Graduate Research Fellowship.

1 Introduction

Since the 1960s cognitive psychologists have used simple laboratory experiments, such as ball-and-urn problems, to show that people do not always process information as perfect Bayesians. For example, participants may under- or over-react to signals depending on how informative they are (Massey and Wu, 2005). Such results speak clearly to our *cognitive* limits.[1] More recently, however, a separate literature has raised the possibility that information processing may be further distorted by *motivated* reasoning. In particular, a person's beliefs about his own (positive) characteristics—which we refer to as "ego"—may affect his utility independent of any effects they have on his actions. This motive may in turn bias information processing, with consequences such as overconfidence in one's own abilities. Overconfidence is thought to have significant economic costs including excess entrepreneurial entry, excess trading, firm overvaluation, and other anomalies in financial markets, as well as excess investment in projects and overpayment for mergers by CEO's.[2]

Motivated by these observations, a fast-growing literature in behavioral economics has explored how exactly people might acquire or process information to "manage" their self-confidence, whether consciously or subconsciously. Some models emphasize selective attention and recall (Rabin and Schrag, 1999; Benabou and Tirole, 2002) while other approaches focus on selective interpretation of information (Akerlof and Dickens, 1982; Brunnermeier and Parker, 2005). Different mechanisms in turn imply different policies for providing feedback. For example, people who tend to "forget" negative feedback might benefit from policies that make it more salient. People who tend to downplay the informativeness of negative feedback, on the other hand, might benefit more from relatively precise, unambiguous feedback. Quarterly performance reviews in companies, for example, might benefit them less than infrequent but more dispositive evaluations.[3]

Identifying the mechanics of self-confidence management requires data on how beliefs evolve over time, and in response to feedback. Cross-sectional data, such as the oft-cited finding that 88% of US drivers rated themselves safer than the median driver (Svenson, 1981), certainly strongly suggest that people are more confident than a Bayesian exposed to the same information would be.[4] But their diagnostic power for identifying *how* people manage their self-confidence is quite limited. For example, Zábojník (2004) and Benoit and Dubra (2011) have shown that purely Bayesian updating can in fact generate highly skewed belief distributions that seem "over-confident."[5]

[1] See Fischhoff and Beyth-Marom (1983) for a review of early work and Benjamin (2019) for a recent review.

[2] See Camerer and Lovallo (1999) and references therein, Odean (1998), Barber and Odean (2001), Daniel et al. (1998), Malmendier and Tate (2005), and Adebambo and Yan (2016). A smaller literature points out that individual overconfidence may nevertheless have aggregate benefits (e.g. Galasso and Simcoe, 2011; Li et al., 2017). Note that in the finance literature the term "overconfidence" is typically used to describe an agent who is too confident in the *precision* of her beliefs about the returns on an asset De Long et al. (1991); we view this as consistent with our notion of overconfidence in one's abilities, in the sense that a forecaster who over-rates their own forecasting abilities will have too little subjective uncertainty about the future as a result.

[3] US firms have recently been shifting from annual to quarterly performance reviews (Church et al., 2012).

[4] Svenson (1981), Englmaier (2006) and Benoit and Dubra (2011) review evidence on overconfidence.

[5] Other interpretations that are consistent with Bayesian updating have also been proposed: people might disagree on the definition of "high ability" (Santos-Pinto and Sobel, 2005) or tend to (rationally) choose activities for which they over-rate their abilities (Van den Steen, 2004). Burks et al. (2013) address this criticism by studying the joint distribution of beliefs and actual ability, but do not directly measure updating. In a related vein, Van den Steen (2011)

The first aim of this paper is to address these concerns by measuring belief updating in a controlled experimental environment in which we can identify specific departures from Bayes' rule and assess to what extent these departures are motivated as opposed to purely cognitive imperfections. Specifically, we measure the evolution of participants' beliefs about their relative performance on an IQ test. We focus on IQ because it affects labor market success across many occupations and we expect participants to have strong ego motives for feeling "smart". We track participants' beliefs about a binary event—scoring in the top half of performers—which lets us summarize their beliefs in a single number, the subjective probability of this event. This in turn lets us elicit beliefs incentive-compatibly using a probabilistic crossover method: we elicit the value of x for which participants are indifferent between receiving a prize with probability x and receiving the prize if their score is among the top half.[6] To minimize participant confusion in the instructions we both alert them that truth-telling is payoff maximizing and also explain the mechanism using narrative story-telling vignettes, alleviating some of the instruction comprehension concerns recently discussed in Danz et al. (2020). We elicit beliefs after the quiz and then repeatedly after providing participants with informative but noisy feedback in the form of signals indicating whether they scored in the top half, which are correct with 75% probability. We then compare belief updates in response to these signals to the Bayesian benchmark to identify which of its properties hold.

We first document three basic properties of Bayes' rule that hold fairly well in our data. First, belief revisions are *invariant* in the sense that the change in beliefs depends only on the information received. One corollary is that we do not find evidence of confirmatory bias, in the sense of participants over-weighting information that confirms their prior views.[7] Second, priors are *sufficient statistics* for past signals with respect to their posteriors, fully summarizing what participants have learned. Third, updating parameters are for the most part stable across rounds of feedback, with no obvious trend. At least in an aggregate sense, then, our data are consistent with some of the basic structural features of Bayes rule.

At the same time, we find that participants do exhibit two substantial biases when incorporating new information into their beliefs. First, they are *asymmetric*, revising their beliefs by 15% more on average in response to positive feedback than to negative feedback. This pattern is consistent across rounds and significant in our preferred specification that pools data across rounds ($p < 0.001\%$), though we are not powered to detect it in all rounds individually. Strikingly, participants who received two positive and two negative signals — and thus learned nothing — ended up significantly

shows how Bayesian-rational agents with differing priors may become overconfident in the sense of over-estimating the precision of their estimates.

[6] Unlike the quadratic scoring rule, this mechanism is robust to risk aversion (and even to non-standard preferences provided participants prefer a higher chance of winning a fixed prize). Allen (1987) and Grether (1992) were among the first, to our knowledge, to discuss this mechanism, but at the time we conducted our experiment it was still fairly novel. In the decade after we conducted our study it has become fairly standard in work on motivated reasoning (see Benjamin, 2019). Experimental economists have widely adopted incentive compatible belief elicitation mechanisms more generally, with much work done exploring both the theoretical properties of these mechanisms and comprehension of underlying incentives by the participants in experiments. See Hossain and Okui (2013), Holt and Smith (2016), and Wilson and Vespa (2017), among others.

[7] See Rabin and Schrag (1999) for a review of work on this topic.

more confident than they began ($p < 0.001\%$). It seems hard to argue that asymmetric updating is the result of a cognitive error rather than a motivated bias in belief formation.

Second, participants are *conservative*, revising their beliefs by only 35% as much on average as unbiased Bayesians with the same priors would. While this could be a motivated bias, conservatism has not previously been linked to self-confidence management.[8] Alternatively, this result could be a simple cognitive error: participants might misunderstand probabilities and treat a "75% correct" signal as less informative than it is. We conduct two tests to distinguish between these two hypotheses about conservatism. First, we show that agents who score well on our IQ quiz—and hence are arguably cognitively more able—are as conservative (and asymmetric) as those who score poorly. Second, we conduct a control experiment, structurally identical to our initial experiment except that participants report beliefs about the performance of a "robot" rather than their own performance. Belief updating in this second experiment is significantly and substantially closer to unbiased Bayesian; in particular, participants are far less conservative. These results suggest that conservatism is at least partly a motivated phenomenon, and that conservatism and asymmetry may be interrelated techniques for managing self-confidence.

Our second aim in the paper is to examine whether the biased beliefs produced by updating processes like this actually matter for subsequent economic decision-making. This is a maintained assumption in the new behavioral literature, which typically models agents who decide whether to compete, invest, etc. by maximizing their expected utility, taking expectations with respect to their biased beliefs. One might wonder, however, whether agents capable of enough "cognitive dissonance" to bias their beliefs initially might also be capable of avoiding such mistakes. They might, for example, maintain one set of beliefs for "consumption" and another for decision-making. To the best of our knowledge there has been no evidence to date on this central issue.

To examine the consequences of biased updating we conduct a second experiment, building on the first. We invite participants to a follow-up in which they perform a real-effort task and must choose between two payment schemes: a piece-rate, and a competitive tournament in which only the highest-performing worker is paid (as in Niederle and Vesterlund, 2007). We again use crossover techniques to measure participants' belief that they will win the tournament, conditional on entering. We show that confidence is *correlated* with competitive behavior, confirming results in previous studies, and then go further to show they have a *causal* effect. To establish causality we exploit our experiment's two-stage structure, using random variation in beliefs induced by feedback in the first experiment to instrument for confidence in the follow-up experiment. We estimate that the effect of confidence on competition is significant and roughly double the magnitude of the OLS correlation. Most importantly, this causal effect of beliefs is stable across more and less conservative

[8] As alluded to above, psychologists tested Bayes' rule for ego-independent problems during the 1960s; conservatism was a common finding. See Slovic and Lichtenstein (1971), Fischhoff and Beyth-Marom (1983), and Rabin (1998) for reviews. See also Grether (1980), Grether (1992) and El-Gamal and Grether (1995) testing whether agents use the "representativeness heuristic" proposed by Kahneman and Tversky (1973). Charness and Levin (2005) test for reinforcement learning and the role of affect using revealed preference data to draw inferences about how participants update. Rabin and Schrag (1999) and Rabin (2002) study the theoretical implications of specific cognitive forecasting and updating biases.

updaters: participants whose beliefs are less responsive to information do not "undo" this bias by making their actions more sensitive to beliefs. More conservative updaters are also less accurate when assessing whether or not they would win a competition. All told, the data suggest that beliefs do affect behavior and that participants who form biased beliefs do not "undo" these biases when choosing their behavior.

Our third aim in the paper is to examine whether the main empirical results—asymmetric and conservative updating of beliefs which in turn influence economic behavior—can be rationalized within a coherent theoretical framework. To this end we build a model of optimal self-confidence management. We discipline the exercise by requiring the agent to process information using rules that match the three properties of Bayes' rule—invariance, sufficiency, and stability—that hold in our (aggregate) data. We call an agent that satisfies these three properties a *biased Bayesian*, as her updating rule satisfies some properties of Bayes' rule but can also accommodate asymmetric or conservative updating.

We study an agent learning about her own ability while balancing rewards of two types: *instrumental utility* from making an investment decision that pays off only if her type is high, and direct *belief utility* from thinking she is a high type, i.e. from her ego. This belief utility can also be interpreted as a reduced-form representation of any number of instrumental reasons for valuing self-confidence. The tension between instrumental and belief utility gives rise to an intuitive first-best: if the agent is of high ability then she would like to learn her type, while if she is a low type she would like to maintain an intermediate belief which trades off ego against accurate decision-making. For example, a mediocre driver might want to think of herself as likely to be a great driver, but not so likely that she drops her car insurance. We then derive the optimal updating bias of an agent who does not know her type and show that it can essentially replicate the agent's first best. Interestingly, the optimal solution requires *both* asymmetry *and* conservatism as natural complementary biases. The intuition is as follows: asymmetry increases the agent's mean belief in her ability in the low state of the world but also increases the variance of the low-type's beliefs, and thus the likelihood of costly investment mistakes. By also updating conservatively the agent can reduce the variance of her belief distribution in the low state of the world.

Our paper contributes in a number of ways to work on motivated reasoning. First, it provides experimental support for two core tenets of behavioral theory on the topic: that people hold positively biased beliefs about their own abilities, and that these beliefs causally affect subsequent decisions. These ideas are common to a wide class of models that examine different *motives* for elevating one's self-confidence: to simply feel good about oneself (Akerlof and Dickens, 1982; Kőszegi, 2006), to derive higher anticipatory utility by believing the future will be bright (Caplin and Leahy, 2001; Brunnermeier and Parker, 2005), to compensate for limited self-control (Brocas and Carrillo, 2000; Carrillo and Mariotti, 2000; Benabou and Tirole, 2002), or to directly enhance performance (Compte and Postlewaite, 2004).

Second, it provides support specifically to theories in which the *mechanism* by which agents manage their self-confidence is through manipulating their beliefs (Akerlof and Dickens, 1982;

Mayraz, 2019) or their interpretations of signals Brunnermeier and Parker (2005).[9] We also find some evidence for a second mechanism, selective acquisition of information, as for example in Carrillo and Mariotti (2000) and Kőszegi (2006); for brevity we discuss these results in Appendix S-2. Our experiment does not speak as directly to a third mechanism, imperfect memory, as it was intentionally designed to minimize forgetfulness (compressing updating into a short time period and reminding participants of the full history of signals at each update).[10]

Third, it contributes to work on "attribution bias," or the tendency to take credit for good outcomes and deny blame for bad ones. One can interpret asymmetric updating in our experiment as an example: participants appear more likely to attribute positive signals to performance and negative signals to noise. This example is not subject to the critique of many earlier studies in social psychology that they "seem readily interpreted in information-processing terms" (Miller and Ross, 1975, p. 224) either because the data-generating processes were not clearly defined,[11] or because key outcome variables were not objectively defined or elicited incentive-compatibly.[12] We address these critiques by examining the evolution of beliefs about a well-defined probabilistic event in response to signals from a clearly specified data-generating process.

Since our results were first released (Mobius et al., 2011) experimental work on belief updating has blossomed, with a number of studies employing (and extending) our and similar designs. Among experimental studies of ego-related beliefs, most of them find support for conservatism while the evidence on asymmetric updating is more mixed.[13] The most closely related work is by Buser et al. (2018) who replicate both our feedback design and the follow-up entry-into-competition experiment. They provide 18 instead of 4 rounds of feedback which allows them to estimate each individual participant's conservatism and asymmetry precisely. There is substantial heterogeneity in both of these traits and conservatism predicts entry into competition. In an intriguing recent paper, Drobner (forthcoming) shows that asymmetry is observed only when subjects do not expect that uncertainty about their performance will be resolved immediately. This is commonly the case in psychology experiments (where there is very strong evidence for asymmetry) but less common

[9]The concept of a single, monolithic "belief" is itself questioned in work on ambiguity aversion, in which agents hold multiple priors and use the most pessimistic to assess any given situation (Gilboa and Schmeidler, 1989). Generally speaking, such models can generate updating patterns inconsistent with Bayes' rule. Because we study a binary event, however, Bayes rule should be satisfied in our data even if participants are ambiguity averse, in the sense that the most pessimistic among a family of priors should yield the most pessimistic among the resulting posteriors.

[10]See Mullainathan (2002), Benabou and Tirole (2002), Wilson (2003), and Gennaioli and Shleifer (2010) for examples of models of imperfect memory.

[11]See Wetzel (1982). For example, in a typical experimental paradigm, participants would teach a student and then attribute the student's subsequent performance either to their teaching or to other factors. A common finding is that participants attribute poor performances to lack of student effort, while taking credit for good performances. This is consistent with the fixed beliefs that (a) student effort and teacher ability are complementary and (b) the teacher is capable.

[12]For example, Wolosin et al. (1973) had participants place 100 metal washers on three wooden dowels according to the degree to which they felt that they, their partner, and the situation were "responsible" for the outcome. Santos-Pinto and Sobel (2005) show that if agents disagree over the interpretation of concepts like "responsibility," this can generate positive self-image on average, and conclude that "there is a parsimonious way to organize the findings that does not depend on assuming that individuals process information irrationally..." (p. 1387).

[13]See Eil and Rao (2011), Grossman and Owens (2012), Charness et al. (2018),Buser et al. (2018) and Coutts (2019), as well as the discussion in Benjamin (2019).

or ambiguous in economic experiments where subjects can often find out how well they did.[14]

Recent work has also extended the analysis in new directions. Several studies examine belief updating about some financial asset in which the participant has a stake (as opposed to own ability) and find systematic deviations from Bayes' rule (though not asymmetric updating) in that domain (Ertac, 2011; Gotthard-Real, 2017; Barron, 2021). Turning to the benefits of self-confidence, Schwardmann and van der Weele (2019) document causal effects on payoffs in a persuasion setting using an identification strategy like the one we use here to study competitive behavior. Finally, Oprea and Yuksel (2021) and Kogan et al. (2021) extend the analysis into social domains, demonstrating asymmetric updating about collective outcomes. Oprea and Yuksel (2021) in particular show that subjects update asymmetrically in response to learning each other's *beliefs*, thus demonstrating the role of self-confidence management in social learning processes.

2 Experimental Design and Methodology

2.1 Quiz Stage

During the *quiz stage*, each participant had four minutes to answer as many questions as possible out of 30. We randomly assigned each participant to one of 9 different versions of the IQ test, and informed participants that quiz types varied and that their performance would be compared against that of students taking the same version. Tests consisted of standard logic and vocabulary questions such as:

> *Question: Which one of the five choices makes the best comparison? LIVED is to DEVIL as 6323 is to (i) 2336, (ii) 6232, (iii) 3236, (iv) 3326, or (v) 6332.*
>
> *Question: A fallacious argument is (i) disturbing, (ii) valid, (iii) false, or (iv) necessary?*

A participant's final score was the number of correct answers minus the number of incorrect answers.[15] Earnings for the quiz were the score multiplied by \$0.25. During the same period an unrelated experiment on social learning was conducted and the combined earnings of all parts of all experiments were transferred to participants' university debit cards at the end of the study. Since earnings were variable and not itemized (and even differed across IQ tests), it would have been very difficult for participants to infer their relative performance from earnings.[16]

Types. We focus on participants' learning about whether or not they scored above the median for their particular IQ quiz. Because these "types" are binary, a participant's belief about her type

[14]Incidentally, it was difficult for subjects to infer their performance in our experiment because we paid them jointly for different parts of the study. This might be one of the reasons why our subjects displayed relatively strong asymmetry in updating.

[15]Quiz questions cover a range of cognitive skills as defined in Cattell-Horn-Carroll theory and in the Woodcock-Johnson test. For example, the pattern recognition question above tests for fluid intelligence (Gf).

[16]Subjects total *expected* earnings from this experiment, given their responses and summing over the quiz stage as well as the feedback and information purchasing stages (described below), had a mean of \$3.94 with standard deviation \$1.88.

at any point in time is given by a single number, her subjective probability of being a high type. This will prove crucial when devising incentives to elicit beliefs, and distinguishes our work from approaches that elicit only several moments of more complicated belief distributions.[17]

2.2 Feedback Stage

During the *feedback stage* we repeated the following protocol four times. First, each participant received a binary signal that indicated whether the participant was among the top half of performers and was correct with 75% probability. We then measured each participant's belief about being among the top half of performers.

Signal Accuracy. Signals were independent and correct with probability 75%: if a participant was among the top half of performers, she would get a "Top" signal with probability 0.75 and a "Bottom" signal with probability 0.25. If a participant was among the bottom half of performers, she would get a Top signal with probability 0.25 and a Bottom signal with probability 0.75. Because the experiment was conducted over the web, we provided a narrative to help participants understand the accuracy of signals. Participants were told that the report on their performance would be retrieved by one of two "robots" — "Wise Bob" or "Joke Bob." Each was equally likely to be chosen. Wise Bob would correctly report Top or Bottom. Joke Bob would return a random report using Top or Bottom with equal probability. We explained that this implied that the resulting report would be correct with 75% probability.

Belief elicitation. To elicit beliefs we use a *crossover* mechanism. Participants were presented with two options,

1. Receive \$3 if their score was among the top half of scores (for their quiz version).

2. Receive \$3 with probability $x \in \{0, 0.01, 0.02, ..., 0.99, 1\}$.

and asked for what value of x they would be indifferent between them. We then draw a random number $y \in \{0, 0.01, 0.02, ..., 0.99, 1\}$ and pay participants \$3 with probability y when $y > x$ and otherwise pay them \$3 if their own score was among the top half. We also told participants that we would elicit beliefs several times but would implement only one choice at random for payment. When we conducted our experiment (April 2005) it was one of the first to use this type of random binary choice mechanism (Healy, 2018).[18]

To explain this mechanism to participants we use a simple but innovative narrative form. We told participants that they were paired with a "robot" who had a fixed but unknown probability y between 0 and 100% of scoring among the top half of participants. Participants could base their chance of winning \$3 on either their own performance or their robot's, and had to indicate the threshold level of x above which they preferred to use the robot's performance. We explained

[17] For example, Niederle and Vesterlund (2007) elicit the mode of beliefs about rank in groups of 4.

[18] Our mechanism is equivalent to Grether's (1992) BDM probability pricing procedure. It has also been independently proposed by Karni (2009) and Holt (1986). Hoelzl and Rustichini (2005) use a related probabilistic mechanism to elicit an indicator for whether participants' beliefs in an event were above or below 1/2.

that participants would maximize their probability of earning \$3 by choosing their own subjective probability of being in the top half as the threshold. Using this narrative device, we thus framed their choice similar to a multiple price list (MPL) which made it easier to explain the procedure to participants in an online experiment.[19]

The crossover mechanism has two main advantages over the quadratic scoring rule for our purposes. First, quadratic scoring is truth-inducing only for risk-neutral participants, while the crossover mechanism is strictly incentive-compatible provided only that participants' preferences are monotone in the sense that among lotteries that pay \$3 with probability q and \$0 with probability $1-q$, they strictly prefer those with higher q.[20] This property holds for von-Neumann-Morgenstern preferences as well as for many non-standard models such as Prospect Theory. Second, the crossover mechanism does not generate perverse incentives to hedge quiz performance. Consider a participant who has predicted she will score in the top half with probability $\hat{\mu}$. Let S denote her score and F her subjective beliefs about the median score $\overline{S}$. Under quadratic scoring she will earn a piece rate of \$0.25 per point she scores and lose an amount proportional to $(I_{S\geq\overline{S}} - \hat{\mu})^2$, so her expected payoff as a function of S is

$$\$0.25 \cdot S - k \cdot \int_{\overline{S}} (I_{S\geq\overline{S}} - \hat{\mu})^2 dF(\overline{S}) \tag{1}$$

for some $k > 0$. For low values of $\hat{\mu}$ this may be *decreasing* in S, generating incentives to "hedge." In contrast, her expected payoff under the crossover mechanism is

$$\$0.25 \cdot S + \$3.00 \cdot \hat{\mu} \cdot \int_{\overline{S}} I_{S\geq\overline{S}} dF(\overline{S}), \tag{2}$$

which unambiguously increases with S. Intuitively, conditional on her own performance being the relevant one (which happens with probability $\hat{\mu}$), she always wants to do the best she can.

2.3 Information Purchasing Stage

In the *information purchasing stage* participants had an opportunity to purchase information about their performance. Participants stated their willingness to pay for receiving \$2 as well as for receiving \$2 and an email containing information on their performance. We bounded responses between \$0.00 and \$4.00. We offered two kinds of information: participants could learn whether they scored in the top half, or learn their exact quantile in the score distribution. For each participant one of these choices was randomly selected and the participant purchased the corresponding bundle if and only if their reservation price exceeded a randomly generated price. This design is a standard

[19] Holt and Smith (2016) demonstrate that "random binary choice" (RBC) methods outperform the quadratic scoring rule in the lab. Hossain and Okui (2013) introduce binarized scoring rules, a generalization of RBC methods for general probability distributions. However, there are concerns that this generality comes at the cost of comprehension (Wilson and Vespa, 2017; Danz et al., 2020; Healy and Kagel, 2021).

[20] See Offerman et al. (2009) for an overview of the risk problem for scoring rules and a proposed risk-correction. One can of course eliminate distortions entirely by not paying participants, but unpaid participants tend to report inaccurate and incoherent beliefs (Grether, 1992).

application of the Becker-DeGroot-Marschak mechanism except that we measure information values by netting out participants' valuations for $2 alone from their other valuations to address the concern that participants may under-bid for objective-value prizes.[21] For the sake of brevity we describe results from this stage in Appendix S-2.

2.4 Follow-up Stage

We invited a random subsample of participants by email one month later for a *follow-up* which repeated the feedback stage but with reference to the performance of a robot, rather than the participant's own performance. participants were told they had been paired with a robot who had a probability θ of being a high type. We then gave participants repeated binary signals of the robot's ability and tracked their beliefs about the robot, just as in the main experiment. To make this comparison as effective as possible we matched experimental conditions in the follow-up as closely as possible to those in the baseline. We set the robot's initial probability of being a high type, θ, to the multiple of 5% closest to the participant's post-IQ quiz confidence. For example, if the participant had reported a confidence level of 63% after the quiz we would pair the participant with a robot that was a high type with probability $\theta = 65\%$. We then randomly picked a high or low type robot for each participant with probability θ. If the type of the robot matched the participant's type in the earlier experiment then we generated the same sequence of signals for the robot. If the types were different, we chose a new sequence of signals. In either case, signals were correctly distributed conditional on the robot's type.

2.5 Competition Stage

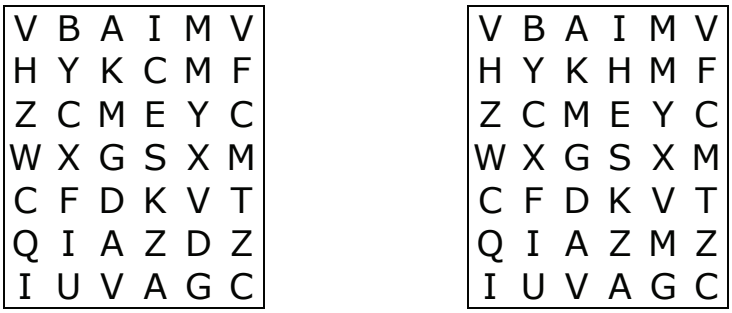
Finally, we invited a second random subsample of participants by email one month after the feedback stage to a *competition stage* based on Niederle and Vesterlund (2007). The purpose of this stage was to analyze the link between self-confidence and willingness to compete in a real effort task. We chose a widely used tournament design in order to proxy for labor market competitions in which self-confidence is important, such as applying for a competitive degree program or selecting into an ambitious career path.

Each participant performed a "character game" multiple times. Each character game consisted of a sequence of character matrices displayed two at a time on screen. Each pair of matrices was identical except for two characters, and the participant's task was to identify the two mismatched characters. Participants could not move on to the next screen without correctly identifying the mismatched characters, and were given three minutes to solve as many puzzles as possible. Participants were given three practice problems before playing for stakes.

Each of the character games was played under distinct incentive conditions, and participants were told that the results from exactly one would be chosen at random and implemented. In the first game all participants were paid on a tournament basis: they were randomly assigned to a

[21] In our data, for example, we find that 89% of participants bid less than $2 for $2. See Appendix S-2.

Figure 1: Sample character game



group of four participants and paid 100 cents per problem solved if they solved the most problems within this group, and nothing otherwise. Before the second game participants were offered a choice: a piece rate of 25 cents per problem solved, or a tournament rate of 100 cents per problem solved if and only if their performance was better than that of the three peers *from the first game*. Participants did not know at this juncture whether they had won the first game. Regardless of their choice, we also separately and subsequently asked each participant what they thought the probability was that they would win such a tournament, i.e. that their score in the second game was higher than that of the others in their group in the first game. We elicited this subjective probability using the same crossover mechanism as above.

3 Data

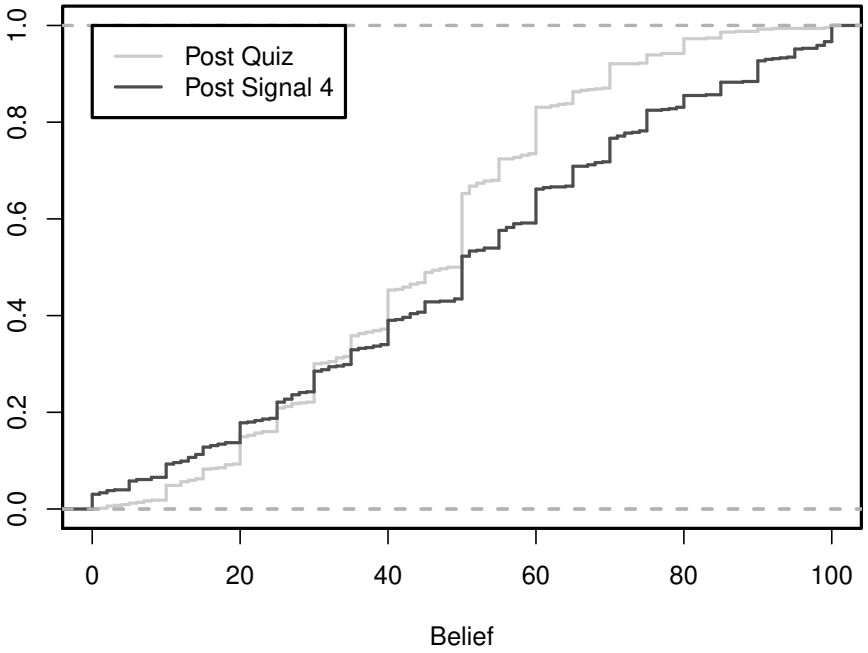
3.1 Participant Pool

The experiment was conducted in April 2005 as part of a larger sequence of experiments at a large private university with an undergraduate student body of around 6,400. A total of 2,356 students signed up in November 2004 to participate in this series of experiments by clicking a link on their homepage on `www.facebook.com`.[22] These students were invited by email to participate in the belief updating study, and 1,058 of them accepted the invitation and completed the experiment online. The resulting sample is 45% male and distributed across academic years as follows: 26% seniors, 28% juniors, 30% sophomores, and 17% freshmen. Our sample includes about 33% of all sophomores, juniors, and seniors enrolled during the 2004–2005 academic year, and is thus likely to be unusually representative of the student body as a whole.

As with any online experiment it is important to consider how well participants understood and followed instructions. Anticipating this issue, our software required participants to make an active choice each time they submitted a belief and allowed them to report beliefs clearly inconsistent with Bayesian updating, such as updates in the wrong direction or simply not updating. After

[22] In November 2004 more than 90% of students were members of the site and at least 60% of members logged into the site daily.

Figure 2: Belief Distributions



Empirical CDFs of participants' beliefs after the quiz (Post Quiz) and after four rounds of feedback (Post Signal 4).

each of the 4 signals, a stable proportion of about 36% of participants reported the same belief as in the previous round.[23] About 16% of participants did not change their beliefs at all during all four rounds of the feedback stage. In contrast, the share of participants who updated in the wrong direction declined over time (13%, 9%, 8% and 7%), and most participants made at most one such mistake.[24] Our preferred estimates use the restricted sample of participants who made no updates in the wrong direction and revised their beliefs at least once. These restrictions exclude 25% and 13% of our sample, respectively, and leave us with 342 women and 314 men. While they potentially bias us against rejecting Bayes' rule, and in particular against finding evidence of conservatism, we implement them to ensure that our results are not driven by participants who misunderstood or ignored the instructions. Our main conclusions hold on the full sample as well and we provide those estimates as robustness checks where appropriate.

To preview overall updating patterns, Figure 2 plots the empirical cumulative distribution function of participants' beliefs both directly after the quiz and after four rounds of updating. Updating yields a flatter distribution as mass shifts towards 0 (for low types) and 1 (for high types). Note that the distribution of beliefs is reasonably smooth, with limited bunching around focal numbers.[25]

[23]The exact proportions were 36%, 39%, 37% and 36% for the four rounds, respectively. Wiswall and Zafar (2014) report similar rates in a feedback experiment where 24% of participants do not update their beliefs about future earnings after being shown the earnings distribution of the corresponding population.

[24]Overall, 19% of participants made only one mistake, 6% made two mistake, 2% made 3 mistakes and 0.4% made 4 mistakes.

[25]Hollard et al. (2016) use this as an approximate metric of participant attentiveness to the decision problem at hand. They compare beliefs elicited using several different procedures and find that the crossover procedure yields

We invited 120 participants to the follow-up stage, of whom 78 participated. The pattern of wrong and neutral moves was similar to the first stage of the experiment. Slightly fewer participants made neutral updates (28% of all updates) and 10% always made neutral updates. Slightly more participants made wrong updates (22% made one mistake, 10% made two mistakes, 5% made three mistakes and 3% made 4 mistakes). The restricted sample for the follow-up has 40 participants.

We invited 274 participants to the competition stage, of whom 146 participated.

3.2 Quiz Scores

The mean score of the 656 participants was 7.4 (s.d. 4.8), generated by 10.2 (s.d. 4.3) correct answers and 2.7 (s.d. 2.1) incorrect answers. The distribution of quiz scores (number of correct answers minus number of incorrect answers) is approximately normal, with a handful of outliers who appear to have guessed randomly. The most questions answered by a participant was 29, so the 30-question limit did not induce bunching at the top of the distribution. Table S-4 in the supplementary appendix provides further descriptive statistics broken down by gender and by quiz type. The 9 versions of the quiz varied substantially in difficulty, with mean scores on the easiest version (#6) fives time higher than on the hardest version (#5). Participants who were randomly assigned to harder quiz versions were significantly less confident that they had scored in the top half after taking the quiz, presumably because they attributed some of their difficulty in solving the quiz to being a low type.[26] We exploit this variation in our data analysis, using quiz assignment as an instrument for beliefs.

4 Information Processing

We now compare empirical belief updating to the Bayesian benchmark. On casual observation they differ starkly: the correlation of participants' logit-beliefs with those predicted by Bayes' rule is 0.57, significantly different from unity. To unpack this result and identify exactly *which* properties of Bayes' rule fail, we next specify empirical models that nest it.

Denote by $\hat{\mu}_t$ the agent's subjective belief after receiving the t^{th} signal, s_t. Considering an arbitrary updating process, we say that it is invariant if it can be written as

$$f(\hat{\mu}_t) - f(\hat{\mu}_{t-1}) = g_t(s_t, s_{t-1}, \ldots) \tag{3}$$

for some sequence of functions g_t that do not depend on $\hat{\mu}_{t-1}$. In particular, invariance excludes *confirmatory bias* where the responsiveness to positive feedback increases with the prior (Rabin and Schrag, 1999). We say that the prior $\hat{\mu}_{t-1}$ is a *sufficient statistic* for information received prior to t if we can write $g_t(s_t, s_{t-1}, \ldots) = g_t(s_t)$. Finally, this relationship is *stable* across time if $g_t = g$ for all t. Together these three properties greatly reduce the potential complexity of information

the smoothest distributions.

[26] Moore and Healy (2008) document a similar pattern.

processing. Bayes' rule satisfies them, as it can be written (in the binary signals case) as

$$\text{logit}(\hat{\mu}_t) = \text{logit}(\hat{\mu}_{t-1}) + I(s_t = H)\lambda_H + I(s_t = L)\lambda_L \tag{4}$$

where $I(s_t = H)$ is an indicator for whether the t^{th} signal was "High", λ_H is the log likelihood ratio of a high signal, and so on. (In our experiment $\lambda_H = -\lambda_L = \ln(3)$.) Our main empirical model nests this Bayesian benchmark:

$$\text{logit}(\hat{\mu}_{it}) = \delta\text{logit}(\hat{\mu}_{i,t-1}) + \beta_H I(s_{it} = H)\lambda_H + \beta_L I(s_{it} = L)\lambda_L + \epsilon_{it} \tag{5}$$

The coefficient δ equals 1 if the invariance property holds, while the coefficients β_H and β_L capture responsiveness to positive and negative information, respectively. The error term ϵ_{it} captures unsystematic errors. Note that we do not include a constant term since $I(s_{it} = H)+I(s_{it} = L) = 1$. To test stability we estimate (5) separately for each of our four rounds of updating and test whether our coefficient estimates vary across rounds. To test whether priors are sufficient statistics we augment the model with indicators $I(s_{i,t-\tau} = H)$ for lagged signals:

$$\begin{aligned}\text{logit}(\hat{\mu}_{it}) = \delta\text{logit}(\hat{\mu}_{i,t-1}) + \beta_H I(s_{it} = H)\lambda_H + \beta_L I(s_{it} = L)\lambda_L \\ + \sum_{\tau=1}^{t-1} \beta_{t-\tau}[I(s_{i,t-\tau} = H)\lambda_H + I(s_{i,t-\tau} = L)\lambda_L] + \epsilon_{it}\end{aligned} \tag{6}$$

Sufficiency predicts that the lagged coefficients $\beta_{t-\tau}$ are zero.

Identifying (5) and (6) is non-trivial because they include lagged logit-beliefs (that is, priors) as a dependent variable. If there is unobserved heterogeneity in participants' responsiveness to information, β_L and β_H, then OLS estimation may yield upwardly biased estimates of δ due to correlation between the lagged logit-beliefs and the unobserved components $\beta_{iL} - \beta_L$ and $\beta_{iH} - \beta_H$ in the error term. Removing individual-level heterogeneity through first-differencing or fixed-effects estimation does not solve this problem but rather introduces a negative bias (Nickell, 1981). In addition to these issues, there may be measurement error in self-reported logit-beliefs because participants make mistakes or are imprecise in recording their beliefs.[27]

To address these issues we exploit the fact that participants' random assignment to different versions of the IQ quiz generated substantial variation in their post-quiz beliefs. This allows us to construct instruments for lagged prior logit-beliefs. For each participant i we calculate the average quiz score of participants *other* than i who took the same quiz variant to obtain a measure of the quiz difficulty level that is not correlated with participant i's own ability but highly correlated with the participant's beliefs.

[27]See Arellano and Honore (2001) for an overview of the issues raised in this paragraph. Instrumental variables techniques have been proposed that use lagged difference as instruments for contemporaneous ones (see, for example, Arellano and Bond (1991)); these instruments would be attractive here since the theory clearly implies that the first lag of beliefs should be a sufficient statistic for the entire preceding sequence of beliefs, but unfortunately higher-order lags have little predictive power when the autocorrelation coefficient δ is close to one, as Bayes' rule predicts.

Table 1: Conservative and Asymmetric Belief Updating

Regressor	Round 1	Round 2	Round 3	Round 4	All Rounds	Unrestricted
Panel A: OLS						
δ	0.814	0.925	0.942	0.987	0.924	0.888
	$(0.030)^{***}$	$(0.015)^{***}$	$(0.023)^{***}$	$(0.022)^{***}$	$(0.011)^{***}$	$(0.014)^{***}$
β_H	0.374	0.295	0.334	0.438	0.370	0.264
	$(0.019)^{***}$	$(0.017)^{***}$	$(0.021)^{***}$	$(0.030)^{***}$	$(0.013)^{***}$	$(0.013)^{***}$
β_L	0.295	0.274	0.303	0.347	0.302	0.211
	$(0.025)^{***}$	$(0.020)^{***}$	$(0.022)^{***}$	$(0.024)^{***}$	$(0.012)^{***}$	$(0.011)^{***}$
$\mathbb{P}(\beta_H = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_L = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_H = \beta_L)$	0.009	0.408	0.305	0.017	0.000	0.000
N	612	612	612	612	2448	3996
R^2	0.803	0.890	0.875	0.859	0.854	0.798
Panel B: IV						
δ	0.955	0.882	1.103	0.924	0.963	0.977
	$(0.134)^{***}$	$(0.086)^{***}$	$(0.122)^{***}$	$(0.123)^{***}$	$(0.058)^{***}$	$(0.060)^{***}$
β_H	0.407	0.294	0.332	0.446	0.371	0.273
	$(0.043)^{***}$	$(0.017)^{***}$	$(0.023)^{***}$	$(0.034)^{***}$	$(0.012)^{***}$	$(0.013)^{***}$
β_L	0.254	0.283	0.273	0.362	0.294	0.174
	$(0.043)^{***}$	$(0.026)^{***}$	$(0.031)^{***}$	$(0.041)^{***}$	$(0.017)^{***}$	$(0.027)^{***}$
$\mathbb{P}(\beta_H = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_L = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_H = \beta_L)$	0.060	0.726	0.086	0.060	0.001	0.004
First Stage F-statistic	13.89	16.15	12.47	12.31	16.48	20.61
N	612	612	612	612	2448	3996
R^2	-	-	-	-	-	-

Notes:

1. Each column in each panel is a regression. The outcome in all regressions is the log posterior odds ratio. δ is the coefficient on the log prior odds ratio; β_H and β_L are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. Bayesian updating corresponds to $\delta = \beta_H = \beta_L = 1$.
2. Estimation samples are restricted to participants whose beliefs were always within $(0, 1)$. Columns 1-5 further restrict to participants who updated their beliefs at least once and never in the wrong direction; Column 6 includes participants violating this condition. Columns 1-4 examine updating in each round separately, while Columns 5-6 pool the 4 rounds of updating.
3. Estimation is via OLS in Panel A and via IV in Panel B, using the average score of other participants who took the same (randomly assigned) quiz variety as an instrument for the log prior odds ratio.
4. Heteroskedasticity-robust standard errors in parenthesis; those in the last two columns are clustered by individual. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.1 Invariance, Sufficiency and Stability

Table 1 presents round-by-round and pooled estimates of Equation 5.[28] Estimates in Panel A are via OLS and those in Panel B are via IV using quiz type indicators as instruments. The F-statistics reported in Panel B indicate that our instrument is strong enough to rule out weak instrument concerns (Stock and Yogo, 2005).

Result 1 (Invariance)**.** *Participants' updating behavior is invariant to their prior.*

Invariance implies that the change in (logit) beliefs should not depend on the prior, or equivalently, that the responsiveness to positive and negative information is not a function of the prior. This implies a coefficient $\delta = 1$ on prior logit-beliefs in Equation 5. The OLS estimate is close to but significantly less than unity; although it climbs by round, we fail to reject equality with one only in Round 4 ($p = 0.57$). These estimates may be biased upward by heterogeneity in the responsiveness coefficients, β_{iL} and β_{iH}, or may be biased downwards if participants report beliefs with noise. The IV estimates suggest that the latter bias is more important: the pooled point estimate of 0.963 is larger and none of the estimates are significantly different from unity.

Of course, it is possible that both β_H and β_L are functions of prior logit-beliefs but that the effects cancel out to give an average estimate of $\delta = 1$. To address this possibility, Table S-7 reports estimates of an augmented version of Equation 5 that includes an interaction between the (logit) prior and the high signal $I(s_{it} = H)$. Invariance requires that the coefficient δ_H on this interaction is zero; our estimated δ_H varies in sign across rounds and is significant at the 5% level only once, in the OLS estimate for Round 1. It is small and insignificant in our pooled estimates using both OLS and by IV, and the same holds when we use our full sample. All told, updating appears invariant at least in the aggregate, though of course this may mask heterogeneity from participant to participant. We note that invariance implies that our participants are *not* prone to confirmatory bias, in the sense that they tend to place more weight on information that is consistent with their priors.

Result 2 (Sufficiency)**.** *Controlling for prior beliefs, lagged information does not significantly predict posterior beliefs.*

Priors appear to be fully incorporated into posteriors—but do they fully capture what participants have learned in the past? Table 2 reports instrumental variables estimates of Equation 6, which includes lagged signals as predictors. We can include one lag in round 2, two lags in round 3, and three lags in round 4. None of the estimated coefficients are statistically or economically significant, supporting the hypothesis that at least in the aggregate priors properly encode past information. We also obtain the same result using the full sample (Table S-5).

Result 3 (Stability)**.** *The structure of updating is largely stable across rounds.*

[28] The logit function is defined only for priors and posteriors in $(0, 1)$; to balance the panel we further restrict the sample to participants i for whom this holds for *all* rounds t. Results using the unbalanced panel, which includes another 101 participant-round observations, are essentially identical.

Table 2: Priors are Sufficient Statistics for Lagged Information

Regressor	Round 2	Round 3	Round 4
δ	0.872	1.124	0.892
	$(0.097)^{***}$	$(0.155)^{***}$	$(0.149)^{***}$
β_H	0.284	0.348	0.398
	$(0.023)^{***}$	$(0.030)^{***}$	$(0.041)^{***}$
β_L	0.284	0.272	0.343
	$(0.027)^{***}$	$(0.031)^{***}$	$(0.028)^{***}$
β_{-1}	0.028	-0.027	0.045
	(0.035)	(0.050)	(0.050)
β_{-2}		-0.036	0.067
		(0.052)	(0.054)
β_{-3}			0.057
			(0.057)
N	612	612	612
R^2	-	-	-

Each column is a regression. The outcome in all regressions is the log posterior odds ratio. Reported coefficients are on the log prior odds ratio (δ), the log likelihood ratio for positive and negative signals (β_H and β_L), and the log likelihood ratio of the signal received τ periods earlier ($\beta_{-\tau}$). The estimation sample includes participants whose beliefs were always within $(0, 1)$ and who updated their beliefs at least once and never in the wrong direction. Estimation is via IV using the average score of other participants who took the same (randomly assigned) quiz as an instrument for the log prior odds ratio. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We test for stability by comparing the coefficients δ, β_H, and β_L across rounds. Our (preferred) IV estimates in Table 1 show some variation, but without any obvious trend. Wald tests for heterogeneous coefficients are mixed; we reject the null of equality for β_H ($p < 0.01$) but not for β_L ($p = 0.24$) or for δ ($p = 0.52$).[29] Given the lack of any clear trends we view the test result for β_H as suggestive, and potentially worth further investigation in a longer panel.

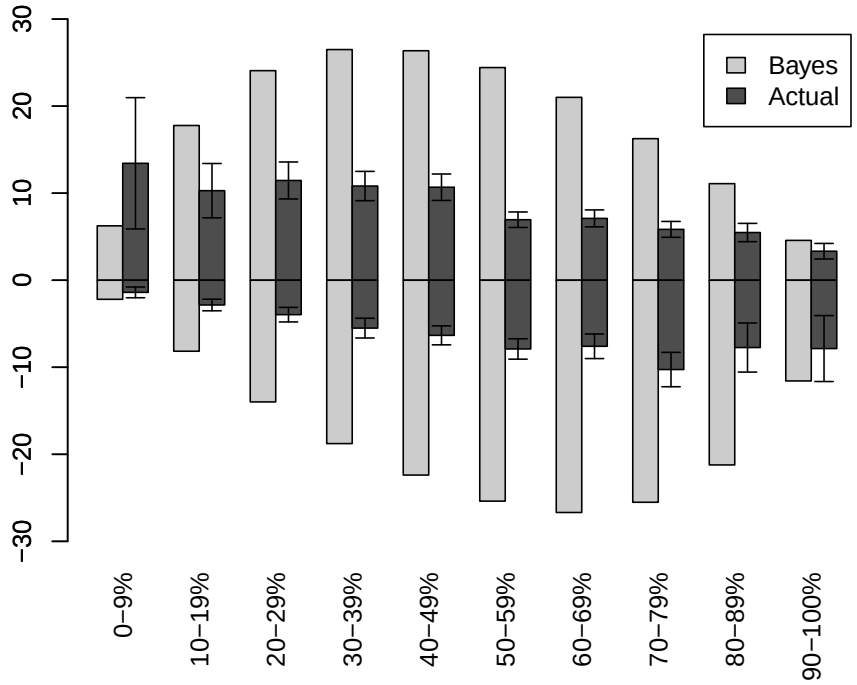
4.2 Conservatism and Asymmetry

Result 4 (Conservatism)**.** *Participants respond less to both positive and negative information than an unbiased Bayesian.*

The OLS estimates of β_H and β_L reported in Table 1, 0.370 and 0.302, are substantially and significantly less than unity. Round-by-round estimates do not follow any obvious trend. The IV and OLS estimates are similar, suggesting there is limited bias in the latter through correlation with lagged prior beliefs. To ensure that this result is not an artifact of functional form, Figure 3 presents a complementary non-parametric analysis of conservatism. The figure plots the mean belief revision in response to a Top and Bottom signal by decile of prior belief in being a top half

[29] We obtain similar results using our full sample, rejecting the null of equality for β_H ($p < 0.01$) but not (quite) for β_L ($p = 0.10$) or for δ ($p = 0.90$).

Figure 3: Conservatism



Mean belief revisions broken down by decile of prior belief in being of type "Top." Responses to positive and negative signals are plotted separately in the top and bottom halves, respectively. The corresponding means that would have been observed if all participants were unbiased Bayesians are provided for comparison. T-bars indicate 95% confidence intervals.

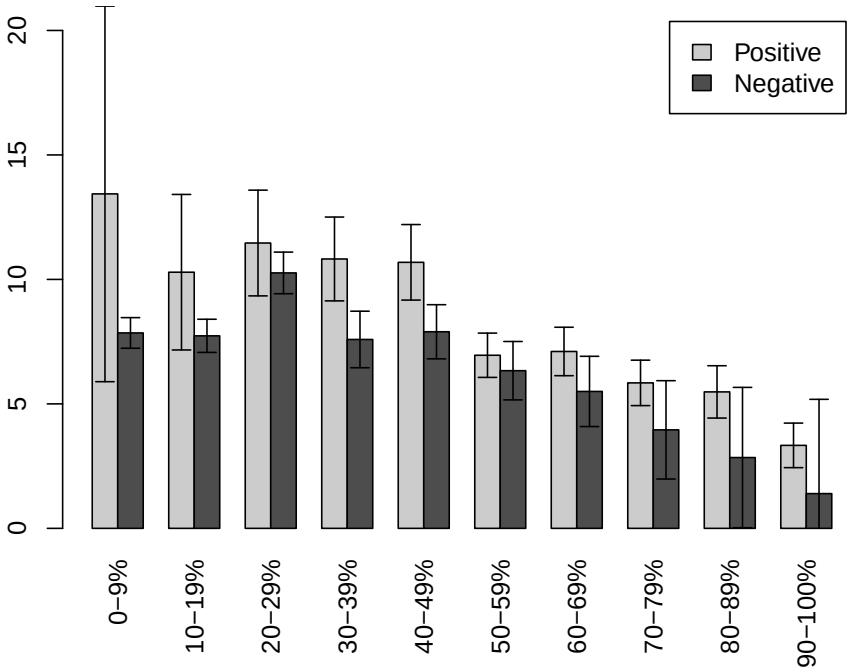
type for each of the four observations of the 656 participants, with the average Bayesian response plotted alongside for comparison. Belief revisions are consistently smaller than those implied by Bayes rule across essentially all of these categories.

Result 5 (Asymmetry)**.** *Controlling for prior beliefs, participants respond more to positive than to negative signals.*

Figure 4 examines this pattern non-parametrically; it compares participants whose prior belief was $\hat{\mu}$ and who received positive feedback with participants whose prior belief was $1 - \hat{\mu}$ and who received negative feedback. According to Bayes' rule, the magnitude of the belief change in these situations should be identical. Instead participants consistently respond more strongly to positive feedback across deciles of the prior. As an alternative non-parametric test we can also examine the net change in beliefs among the 224 participants who received two positive and two negative signals. These participants should have ended with the same beliefs as they began; instead their beliefs increased by an average of 4.8 percentage points ($p < 0.001$).

To quantify asymmetry we turn to our regression analysis, comparing estimates of β_H and β_L, the responsiveness to positive and negative signals, from Table 1. The difference $\beta_H - \beta_L$ is consistently positive across all rounds and significantly different from zero in our preferred (and best-powered) pooled specification. It is also significant in the first and fourth rounds individually, though not the second and third, and we cannot reject the hypothesis that the estimates are equal

Figure 4: Asymmetry



Mean absolute belief revisions by decile of prior belief in being of type equal to the signal received. For example, a participant with prior belief $\hat{\mu} = 0.8$ of being in the top half who received a signal T and a participant with prior belief $\hat{\mu} = 0.2$ who received a signal B are both plotted at $x = 80\%$. T-bars indicate 95% confidence intervals.

across all four rounds ($p = 0.32$). The IV estimates are somewhat more variable but are again uniformly positive, and significantly so in our preferred pooled specification. All told, while we are not powered to detect it in some rounds of updating individually, we view this as strong cumulative evidence of asymmetric updating.

The size of the difference is substantial, implying that the combined effect of receiving both a positive and a negative signal (equivalent to getting no information) is positive and, in terms of magnitude, approximately a quarter (26%) of the effect of receiving only a positive signal.[30]

Deviations from Bayes' rule were costly within the context of the experiment. Comparing expected payoffs given observed updating (π_{actual}) to those participants would have earned if they updated using Bayes' rule (π_{Bayes}) or if they did not update at all ($\pi_{noupdate}$), we find that the ratio $\frac{\pi_{Bayes}-\pi_{actual}}{\pi_{Bayes}-\pi_{noupdate}}$ is 0.59. Non-Bayesian updating behavior thus costs participants 59% of the potential gains from processing information within the experiment.

4.3 Is Conservatism a Cognitive Failing?

Before discussing our results, we first test for a potential confound in our interpretation of conservative updating. Unlike asymmetry, conservatism could potentially be a cognitive rather than

[30]Table S-6 in the supplementary appendix shows that the results of the regression continue to hold when we pool all four rounds of observation, even when we eliminate all observations in which participants do not change their beliefs. That is, the effect is not driven by an effect of simply not updating at all.

Table 3: Heterogeneity in Updating by Ability

Regressor	OLS
β_H	0.426
	$(0.029)^{***}$
β_L	0.281
	$(0.014)^{***}$
β_H^{Able}	-0.072
	$(0.033)^{**}$
β_L^{Able}	0.012
	(0.025)
N	2448
R^2	0.405

Each column is a separate regression. The outcome in all regressions is the log belief ratio. δ, β_H, and β_L are the estimated effects of the prior belief and log likelihood ratio for positive and negative signals, respectively. δ^{Able}, β_H^{Able}, and β_L^{Able} are the differential responses attributable to high ability. Robust standard errors clustered by individual reported in parentheses. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

motivated bias. Participants might misinterpret the informativeness of signals and treat them as if they were correct with probability less than 75%. This would be understandable if participants usually encountered weaker signals in daily life, for example.

We present two pieces of evidence that suggest that cognitive errors are not the driving factor. First, we show that conservatism (and asymmetry) do not correlate with the cognitive ability of participants. Specifically, we assess whether biases are present both among high performers (those that score in the top half) and low performers on the IQ quiz. Table 3 reports estimates of Equation 5 differentiated by ability. We find no evidence that more able (higher performing) participants update differently than less able participants: they do not differ in the way they weight their priors or in the way they incorporate positive and negative signals. Of course, to the extent that our intelligence measure does not perfectly capture information-processing ability this test is only suggestive.

For a more definitive test distinguishing motivated behavior from cognitive errors we turn to the results of the follow-up experiment, in which a random subset of participants performed an updating task that was formally identical to the one in the original experiment, but which dealt with the ability of a robot rather than their own ability. For these participants we pool the updating data from both experiments and estimate:

$$\begin{aligned} \text{logit}(\hat{\mu}_{it}^e) - \text{logit}(\hat{\mu}_{it}^e) = \beta_H \cdot I(s_{it} = H)\lambda_H + \beta_L \cdot I(s_{it} = L)\lambda_L + \\ +\beta_H^{Robot} \cdot I(e = \text{Robot}) \cdot I(s_{it} = H)\lambda_H + \beta_L^{Robot} \cdot I(e = \text{Robot}) \cdot I(s_{it} = L)\lambda_L + \epsilon_i^t \end{aligned} \tag{7}$$

Here, e indexes experiments (Ego or Robot), so that the interaction coefficients β_H^{Robot} and β_L^{Robot} tell us whether participants process identical information differently across both treatments. Given the smaller sample available we impose $\delta = 1$ and estimate by OLS. Table 4 reports results.

Table 4: Belief Updating: Own vs. Robot Performance

Regressor	I	II	III
β_H	0.426	0.349	0.252
	$(0.087)^{***}$	$(0.066)^{***}$	$(0.043)^{***}$
β_L	0.330	0.241	0.161
	$(0.050)^{***}$	$(0.042)^{***}$	$(0.033)^{***}$
β_H^{Robot}	0.362	0.227	0.058
	$(0.155)^{**}$	$(0.116)^{*}$	(0.081)
β_L^{Robot}	0.356	0.236	-0.006
	$(0.120)^{***}$	$(0.085)^{***}$	(0.089)
$\mathbb{P}(\beta_H + \beta_H^{Robot} = 1)$	0.128	0.000	0.000
$\mathbb{P}(\beta_L + \beta_L^{Robot} = 1)$	0.004	0.000	0.000
$\mathbb{P}(\beta_H = \beta_L)$	0.302	0.118	0.039
$\mathbb{P}(\beta_H + \beta_H^{Robot} = \beta_L + \beta_L^{Robot})$	0.454	0.316	0.030
N	160	248	480
R^2	0.567	0.434	0.114

Each column is a separate regression. The outcome in all regressions is the change in the log belief ratio. β_H and β_L are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. β_H^{Robot} and β_L^{Robot} are the differential responses attributable to obtaining a signal about the performance of a robot as opposed to one's own performance. Estimation samples are restricted to participants who participated in the follow-up experiment and observed the same sequence of signals as in the main experiment. Column I includes only participants who updated at least once in the correct direction and never in the wrong direction in both experiments. Column II adds participants who never updated their beliefs. Column III includes all participants. Robust standard errors clustered by individual reported in parentheses. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Result 6. *Conservatism is significantly reduced when participants learn about a robot's performance rather than their own performance.*

The baseline coefficients β_H and β_L are similar to their estimated values for the larger sample (see Table 1), suggesting that participation in the follow-up was not selective on updating traits. The interaction coefficients are both positive and significant—they imply that participants are roughly twice as responsive to feedback when it concerns a robot's performance as they are when it concerns their own performance. In fact, we cannot reject the hypothesis that $\beta_H + \beta_H^{Robot} = 1$ ($p = 0.13$), though we can still reject $\beta_L + \beta_L^{Robot} = 1$ ($p = 0.004$). While conservatism does not entirely vanish, it is clearly much weaker. Interestingly, participants are also less asymmetric in relative terms when they update about robot performance $\left(\frac{\beta_H}{\beta_L} > \frac{\beta_H + \beta_H^{Robot}}{\beta_L + \beta_L^{Robot}}\right)$. We cannot reject the hypothesis that they update symmetrically about robot performance such that $\beta_H + \beta_H^{Robot} = \beta_L + \beta_L^{Robot}$ ($p = 0.45$).[31]

[31]The robot condition might not eliminate all updating biases if these biases are due to anticipatory motives as in Brunnermeier and Parker (2005): participants might still want to believe that their robot will earn them a cash prize. The fact that updating about robots is significantly different from updating about ones' self lets us rule out both purely cognitive interpretations and purely anticipatory ones, but it is possible that anticipation still plays some role. On the other hand, the fact that updating about robots is not significantly asymmetric suggests that we cannot reject the null of no anticipatory utility.

5 Biased Updating and Economic Outcomes

We have analyzed how participants actively manage their self-confidence. However, we have not yet demonstrated that self-confidence management is *economically* significant: unless biased updating leads to biased outcomes we are just describing a psychological phenomenon. The link between biased beliefs and actions is a largely unquestioned assumption in the literature. We believe however that it is the key that turns a psychological phenomenon into an economically relevant one. Basically, we need to address the skepticism that agents who can control how they form beliefs may also control how they use such biased beliefs and thereby potentially neutralize any biases accrued while forming beliefs. We therefore next present results from our second follow-up experiment. As a reminder, participants in this experiment participated in an effort task and chose between two compensation schemes: a piece rate, and a tournament scheme in which only top scorers were paid (see Section 2.5 for full details).

5.1 Do Beliefs Matter?

We first examine whether reported beliefs drive behavior. Letting $A_i^{CS} \in \{0, 1\}$ indicate participant i's decision to compete in the tournament stage, and $\hat{\mu}_i^{CS}$ her reported subjective probability of winning that tournament, we are interested in estimating

$$A_i^{CS} = \varphi + \vartheta \hat{\mu}_i^{CS} + \epsilon_i \tag{8}$$

Column 1 of Table 5 estimates this equation via OLS and finds a positive and significant association: a 1% increase in the subjective probability of winning is associated with a 1% increase in the probability of competing. Earlier work also typically finds a positive correlation between $\hat{\mu}_i^{CS}$ and A_i^{CS} (e.g. Niederle and Vesterlund (2007)). The interpretation of this result is clouded, however, by the fact that beliefs are endogenous variables potentially correlated with any number of other unobservable personal attributes that affect competitiveness.

To address this problem we exploit exogenous variation in beliefs generated by our first experiment (feedback stage) to instrument for self-confidence in the competition stage. The basic idea is simple: conditional on whether a participant actually scored in the top half, the number of positive signals that participant observed is purely random. We therefore use this sum as an instrument for confidence prior to competing in the second experiment.[32] Column 2 of Table 5 shows that relative performance feedback affects beliefs on relative performance among the same pool of participants that persisted across time and across domains. Each positive signal received in the first experiment increases a participant's subjective belief that they can win a competition by 5 percentage points ($p < 0.05$). Besides establishing instrument relevance, this also illustrates that participants took

[32] One potential concern with this strategy is that the number of positive signals a participant observed affected their likelihood of participating in the second experiment, generating selection bias. To test this we regressed an indicator for participation on the number of positive signals received, conditional on ability, and estimated a small ($\beta = 0.01$) and insignificant ($p = 0.74$) relationship. Results available on request.

Table 5: Confidence and Competition: IV Estimates

	OLS	First Stage	IV	Reduced Form	Over-controlled
Confidence (Experiment 2)	0.010		0.024		0.009
	(0.002)***		(0.010)**		(0.002)***
Feedback (Experiment 1)		5.159		0.125	0.077
		(2.122)**		(0.053)**	(0.053)
Ability (Experiment 1)	0.222	-9.488	0.226	-0.004	0.084
	(0.089)**	(5.397)*	(0.091)**	(0.139)	(0.131)
N	102	102	102	102	102
R^2	0.221	0.056	-	0.104	0.240

Notes: Each column reports a separate regression. The outcome in Columns 1 and 3-5 is an indicator equal to 1 if the participant chose to compete in Experiment 2; the outcome in Column 2 is the participant's subjective probability of winning the competition in Experiment 2. The regressors are the participant's subjective probability of winning the competition in Experiment 2 ("Confidence"), the sum of the signals the participant received in Experiment 1 ("Feedback"), and an indicator for whether the participant scored in the top half in Experiment 1 ("Ability"). Estimation via OLS is reported in Columns 1-2 and via instrumental variables in Column 3 using "Feedback" as the excluded instrument. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

information from the first experiment seriously, as it had a lasting impact on their self-confidence four weeks later.

In Column 3 we use this instrument to estimate the causal effect of confidence on competitiveness. We find that confidence has a significant, positive effect on the probability of competing. Moreover, the estimated magnitude of this effect is more than twice as large as the OLS estimate: a 1% increase in the subjective probability of winning causes a 2.4% increase in the probability of competing. This size difference, whether due to measurement error in beliefs or to unobserved heterogeneity, implies that beliefs may have substantially more explanatory power for decisions to compete than earlier OLS estimates imply. More broadly, the result confirms that beliefs about one's own ability matter for subsequent decision-making.

Result 7. *A participant's confidence causally increases her propensity to enter a tournament.*

One outstanding concern is that the beliefs we measure might be far from a sufficient statistic for behavior. For example, participants might maintain dual mental systems, one of which responds to questions like "how likely are you to win" (even when incentivized) while another guides decisions like whether to compete. If this were true then the feedback participants receive concerning their relative performance from our first experiment would affect decision-making in the second above and beyond, or independent of its effect on reported beliefs $\hat{\mu}_i^{CS}$. To test this hypothesis, Columns 4 and 5 examine the reduced-form impact of information from the first experiment on competition in the second before and after controlling for beliefs. While positive feedback does increase competitiveness, this effect shrinks and becomes insignificant once we control for reported beliefs. The data thus support the view that beliefs to a large extent incorporate information accrued from the

Table 6: Confidence affects competition similarly for more and less conservative types

	First Stage		Reduced Form		IV	
	LC	MC	LC	MC	LC	MC
Confidence (Experiment 2)					0.036	0.030
					(0.015)**	(0.027)
Feedback (Experiment 1)	6.640	2.429	0.238	0.073		
	(4.050)	(2.583)	(0.077)***	(0.075)		
Ability (Experiment 1)	-6.126	-12.984	-0.097	0.030	0.123	0.422
	(9.921)	(6.228)**	(0.198)	(0.203)	(0.208)	(0.310)
N	49	47	49	47	49	47
R^2	0.087	0.089	0.272	0.045	-	-

Notes: Each column reports a separate regression. The outcome in Columns 1-4 is the participant's subjective probability of winning the competition in Experiment 2; the outcome in Columns 5-6 is an indicator equal to 1 if the participant chose to compete in Experiment 2. The estimation sample includes participants who update more (less) conservatively than the median in columns labelled "MC" ("LC"). The regressors are the participant's subjective probability of winning the competition in Experiment 2 ("Confidence"), the sum of the signals the participant received in Experiment 1 ("Feedback"), and an indicator for whether the participant scored in the top half in Experiment 1 ("Ability"). Estimation via OLS is reported in Columns 1-4 and via instrumental variables in Columns 5-6 using "Feedback" as the excluded instrument. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

past.

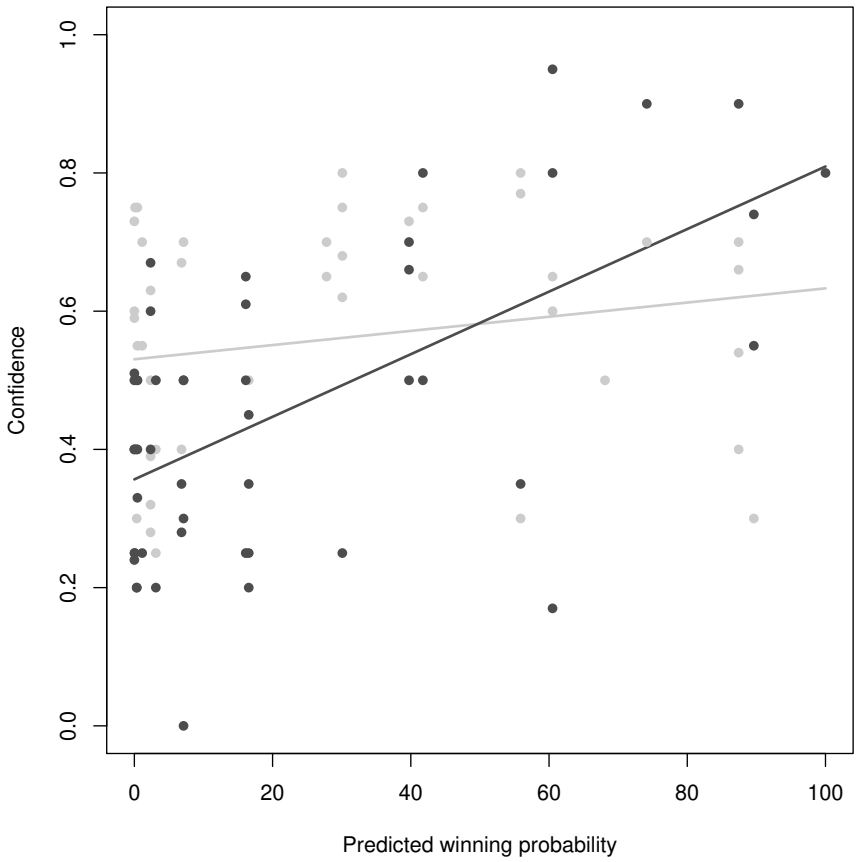
5.2 Can Biased Updating be "Unraveled"?

While we have shown that (biased) beliefs do affect behavior, it is still conceivable that participants can at least partially "undo" the effects of their biased belief updating when time comes to decide. For example, it could be that more conservative agents who have beliefs that are less sensitive to feedback than their peers in turn make decisions that are *more* sensitive to beliefs. On net, these agents' behavior could respond similarly to feedback as their less conservative peers. Conservatism would then amount to a simple "re-scaling" of beliefs rather than a bias that affects behavior.

To test whether conservatism in updating affects behavior we categorize participants into those who update more conservatively (MC) and less conservatively (LC) than average in our base experiment.[33] We then repeat the instrumental variables estimation process reported in Table 5 separately for each of these two groups. If beliefs have a different scale for more conservative updaters we should see larger IV coefficients for the more conservative group. In fact we find the opposite (Table 6); the IV point estimate for more conservative updaters is similar to and slightly smaller than that for less conservative ones. While the former estimate is imprecise, there is no direct evidence that beliefs mean something different for more conservative updaters.

[33] We do this as follows. First, for each round r and signal type we rank participants by the magnitude of the change in their logit belief, and normalize these ranks to $[0, 1]$. Second, we define a participant's overall responsiveness to information as the average across all four rounds of their rank-responsiveness. Third, we define participants as "more responsive" if their average rank exceeds 0.5 and "less responsive" otherwise.

Figure 5: More conservative updaters have less accurate beliefs.



Plots the relationship between subjective and objective probabilities of winning a tournament separately for less conservative (dark grey) and more conservative (light grey) participants.

Result 8. *Confidence affects decisions equally for more and less conservative participants.*

Our data allow us to conduct a second test of the impact of conservatism. Specifically, conservatism implies that participants' beliefs respond less to information. Therefore, more conservative participants should have less self-knowledge than less conservative participants. Figure 5 illustrates the degree of self-knowledge of more- and less-conservative agents. Formally, it plots the relationship between participants' subjective beliefs that they would win a tournament and their objective probability, given their score. The relationship is significantly and substantially weaker for more conservative updaters.

Our two-experiment design allows us to instrument for beliefs (by directly affecting them through feedback) providing direct evidence of the effect of beliefs on behavior. Furthermore, we showed that beliefs affect behavior equally, independently whether such beliefs derived from agents who are more or from agents who are less conservative in updating beliefs. Therefore, biases

that weaken the relationship between types of participants and beliefs in turn weaken the relationship between types and *actions*, leading to mistakes. Specifically, high-ability types who update conservatively will tend to take too few risks, while low-ability types who do so will take too many.

6 Optimally Biased Bayesian Updating

Our experimental data show that participants update their beliefs with substantial biases and that these beliefs then drive subsequent decision-making. At the same time, some properties of Bayesian updating (invariance, sufficiency and stability) do appear to hold quite well in our data, at least in an aggregate sense. In this section we show that these properties provide enough structure to model biased updating coherently in an optimizing framework, and that the biases evident in our data emerge naturally as a result.

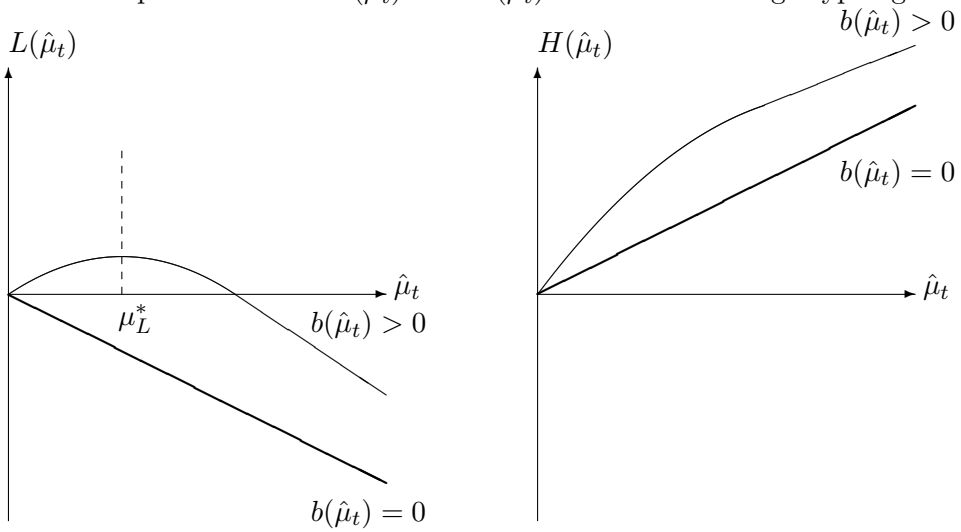
We study an agent who is a high type H with probability μ_0 and otherwise a low type L (reflecting our experimental design). There are T discrete time periods in each of which the agent receives a signal s_t about her ability. The agent aggregates the stream of signals up to time t into a *subjective belief* $\hat{\mu}_t$. We allow the agent's belief to differ from the *objective probability* μ_t derived using Bayes' rule. The agent balances two objectives when forming biased subjective beliefs: she wants to make good instrumental decisions, but also cares about her ego and wants to believe that she is a high type. We first define instrumental and belief utility formally and derive the agent's optimal beliefs if she could choose them freely. We then derive the constrained-optimal updating behavior of biased Bayesians who manage their self-confidence. We also show that this bias remains approximately optimal even if the agent's instrumental and belief utility changes, which lets us think of the optimal bias as an evolutionary adjustment.

6.1 Utility and Optimal Beliefs

We start with instrumental utility. With equal probability, nature selects one of the T time periods as the "investment period". In this period the agent must decide whether or not to take an action that yields a positive payoff if and only if her type is high. For example, the agent might consider investing in the stock market and has to decide if she is a skilled investor, or she might consider taking a challenging major in college and has to decide whether she is smart enough. Formally, the agent can make an investment which pays 1 in the final period T if she is of high type or 0 otherwise.[34] The investment has a cost $c \in [0, 1]$ which is drawn from a well behaved continuous distribution $G \in C^2[0, 1]$ at the time of the decision. Not investing gives a payoff of 0. The optimal decision of a Bayesian decision maker is thus to invest if and only if $c < \mu_t$. Consistent with the results of our second experiment, we assume that a biased agent behaves *as if she were a Bayesian* and invests iff $c < \hat{\mu}_t$. Hence, biasing updating is costly because it leads to worse decisions.

[34]The assumption that the instrumental value of investing is realized in the last period simplifies our calculation of belief utility because the agent only learns her type in the final period and therefore manages her belief utility over all time periods $1 \leq t \leq T$.

Figure 6: Per-period utilities $L(\hat{\mu}_t)$ and $H(\hat{\mu}_t)$ of the low and high type agents



The agent also derives direct *belief utility* $b(\hat{\mu}_t)$ in period t from her subjective belief, where $b \in C^2[0,1]$ is a well-behaved, strictly increasing function normalized such that $b(0) = 0$. The model is agnostic over the various kinds of belief utility discussed in the literature; to capture them in a reduced-form way we make no assumptions about the shape of $b(\cdot)$ other than monotonicity.[35] The combined objective function of the agent is the sum of her average belief utility and her expected instrumental utility:

$$U(\hat{\mu}_0, .., \hat{\mu}_T) = \frac{1}{T} \sum_{t=1}^{T} \left[\underbrace{b(\hat{\mu}_t)}_{\text{belief utility}} + \underbrace{\int_0^{\hat{\mu}_t} (\mu_t - c)\, dG(c)}_{\text{instrumental utility}} \right] \tag{9}$$

When $b(\hat{\mu}) = 0$ the agent has no belief utility and behaves like a classical economic agent. Note that because payoffs are time-averaged T serves as a measure of the information-richness of the environment. In stating results we will make use of the notion of *relative time* $\tau \in [0,1]$ which we associate with absolute time $\lfloor \tau T \rfloor$.

To build intuition it will be useful to study the per-period expected utility of the low and high type agents, which we denote $L(\hat{\mu}_t)$ and $H(\hat{\mu}_t)$:

$$\begin{aligned} L(\hat{\mu}_t) &= b(\hat{\mu}_t) - \int_0^{\hat{\mu}_t} c dG(c) \\ H(\hat{\mu}_t) &= b(\hat{\mu}_t) + \int_0^{\hat{\mu}_t} (1-c) dG(c) \end{aligned} \tag{10}$$

[35] In our model, subjective beliefs will converge for most time periods as $T \to \infty$. Other models in the literature analyze settings with few feedback periods where subjective beliefs remain noisy and hence the concavity or convexity of the belief utility function matters (see for example Kőszegi (2006)).

Suppose for now that agents of low and high type could *choose* subjective beliefs μ_L^* and μ_H^* to maximize these respective expressions. As Figure 6 illustrates, the high type agent would always choose $\mu_H^* = 1$ because both her belief and instrumental utility are increasing in her subjective belief. The optimal (and possibly non-unique) μ_L^* for the low type agent depends on $b(\cdot)$, however: an agent without belief utility chooses $\mu_L^* = 0$ while an agent with ego concerns may choose $\mu_L^* > 0$. We focus on the interesting case $\mu_L^* > 0$ in which the low-type agent prefers on net to hold an inflated belief.[36] We also restrict attention to decision problems with $L(1) < 0$ which implies $\mu_L^* < 1$, or in other words that the low-type agent would not want to convince herself that she was the high type. While this extreme form of bias is conceivable in situations where there are no real stakes (or belief utilities are large), it generates no interesting predictions.

6.2 Optimal Biased Bayesian Updating

Agents receive a stream of i.i.d. signals in each period t. A signal can take finitely many values which we index by k ($1 \leq k \leq K$) with distribution F_H in the high state and F_L in the low state. Let $\lambda_k = \log(F_H(k)/F_L(k))$ be the log-likelihood ratio for realization k. Every signal realization is informative such that $\lambda_k \neq 0$. Motivated by our experimental results, we assume that agents update their belief as biased Bayesians whose updating process satisfies invariance, sufficiency and stability.

Definition 1. *A biased Bayesian updating process consists of an initial subjective prior $\hat{\mu}_0$ and an updating rule*

$$\mathit{logit}(\hat{\mu}_{t+1}) = \mathit{logit}(\hat{\mu}_t) + \beta_k \lambda_k \tag{11}$$

where $\beta_k \geq 0$.

We refer to β as the *responsiveness function* and to $\tilde{\beta}_k = \beta_k / \max_k \beta_k$ as the *normalized responsiveness*.[37] Biased Bayesian updating encompasses standard Bayesian updating as a special case ($\hat{\mu}_0 = \mu_0$ and $\beta_k = 1$) while capturing the idea that the agent may downplay or overstate the informativeness of certain kinds of feedback. Following Brunnermeier and Parker (2005), we say that a biased Bayesian updating process is *optimal* if it maximizes expected total utility (9) among all such processes.[38] We do not take a strong view here on the extent to which the agent need be conscious of biasing her updating; one can also interpret optimal bias as the result of a subconscious tendency to select habits of thought that increase well-being, or of a longer-run evolutionary process.

When the agent has no belief utility the optimum is, reassuringly, to be unbiased.

[36] It is not difficult to come up with conditions such that $\mu_L^* > 0$. For example, any linear belief utility function will suffice. We know that $L(0) = 0$ and $L(1) < 0$. Moreover, for small x we have $L(x) > 0$ because G' is continuous and hence bounded and therefore $\int_0^x c dG(c) \leq \int_0^x c \max_{c \in [0,1]} (G'(c))\, dc = \frac{1}{2} (x)^2 \max_{c \in [0,1]} (G'(c))$.

[37] The normalized responsiveness is only defined for responsiveness functions which are not zero everywhere.

[38] Existence is guaranteed since (a) expected utility is continuous in $\hat{\mu}_0 \in (0,1)$ and β_k; (b) using the logic of proposition 2, one can show that there are $\epsilon > 0$ and $M > 0$ such it is never optimal to choose $\hat{\mu}_0 < \epsilon$, $\hat{\mu}_0 > 1 - \epsilon$ or $\beta_k > M$. Hence, the optimal parameters live in a compact Euclidean metric space.

Proposition 1. *Let $T \geq 2$. The optimal biased Bayesian updating process for an agent without belief utility ($b(\hat{\mu}) = 0$ for all $\hat{\mu}$) is Bayes' rule: $\hat{\mu}_0 = \mu_0$ and $\beta_k = 1$ for all k.*

To characterize the case with belief utility we introduce the notions of *conservatism* and *downward neutral bias*, which is a strong form of *asymmetry*.

Definition 2. *A biased Bayesian updating process is **conservative** if the agent always responds less to new information than an unbiased Bayesian ($\max_k \beta_k < 1$). It exhibits a **downward neutral bias** (DNB) if $\sum_k F_L(k)\tilde{\beta}_k\lambda_k = 0$.*

DNB implies that the agent's expected logit-belief remains unchanged if the state is low; the agent essentially interprets the stream of information as white noise. DNB is a generalized notion of *asymmetry*: in the binary signals case, if H (L) denotes the signal with the higher (lower) log-likelihood ratio, DNB implies $\beta_H > \beta_L$.

Proposition 2. *The optimal updating process has the following features: (1) $\beta_k^T \to 0$ as $T \to \infty$ for all k so that the agent updates conservatively for large T; (2) $\sum_k F_L(k)\tilde{\beta}_k^T\lambda_k \to 0$ as $T \to \infty$ so that the agent exhibits DNB for large T; (3) if moreover the low type's optimal belief μ_L^* is unique and $L''(\mu_L^*) < 0$ then $\hat{\mu}_0^T \to \mu_L^*$; (4) for any relative time $\tau > 0$ the agent's belief converges in probability to μ_L^* in the low state and to $\mu_H^* = 1$ in the high state.*

The intuition for this result can be illustrated graphically for the binary signals case. The evolution of logit-beliefs described in Equation 11 follows a random walk: in each period, the logit-belief increases by $\beta_H\lambda_H$ with probability $F_H(H)$ for the high type ($F_L(H)$ for the low type) and otherwise decreases by $\beta_L\lambda_L$. The mean logit-belief of the high type, $\hat{\gamma}_t^H$, and the variance in logit-beliefs, $\left(\hat{\sigma}_t^H\right)^2$, can hence be expressed as:

$$\begin{aligned} \hat{\gamma}_t^H &= \text{logit}(\hat{\mu}_0) + t\left[F_H(H)\beta_H\lambda_H + (1 - F_H(H))\beta_L\lambda_L\right] \\ \left(\hat{\sigma}_t^H\right)^2 &= tF_H(H)(1 - F_H(H))\left(\beta_H\lambda_H - \beta_L\lambda_L\right)^2 \end{aligned} \tag{12}$$

We can derive analogous expressions $\hat{\gamma}_t^L$ and $\left(\hat{\sigma}_t^L\right)^2$ for the mean and variance of the low type's logit-belief by replacing the probability $F_H(H)$ with $F_L(H)$. The left panel of Figure 7 shows the mean logit belief of the high type (increasing solid line) and low type (decreasing solid line) when the agent is an unbiased Bayesian. Note that the mean logit beliefs of both types converge to $+\infty$ and $-\infty$ at rate t while the standard deviation increases only at rate $\sqrt{t}$. Therefore, beliefs converge to either 1 or 0 in probability.

The biased Bayesian would prefer keep her beliefs close to either 1 (in the high state) and $\mu_L^* > 0$ (in the low state). By choosing an initial belief close to her optimal low-type's belief μ_L^* and by becoming asymmetric ($\beta_H/\beta_L \uparrow$) she can slow the rate at which the low type's logit-belief drifts to $-\infty$, or even eliminate this drift altogether by choosing a DNB. The right panel of Figure 7 illustrates this idea. Asymmetry alone is insufficient, however, without conservatism: unless the agent also reduces her responsiveness to information the variance of the low type's logit-beliefs will

Figure 7: Evolution of logit-beliefs of an unbiased Bayesian (left panel) and an optimally biased Bayesian (right panel)

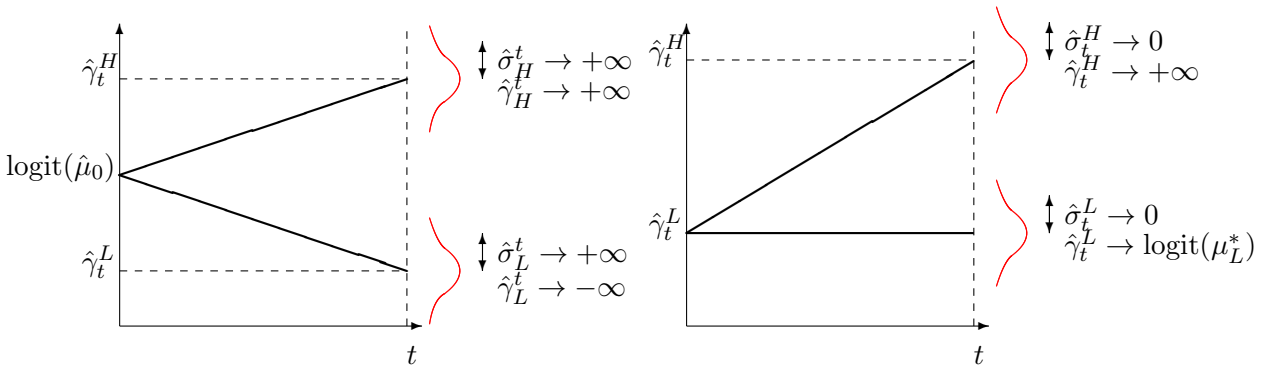
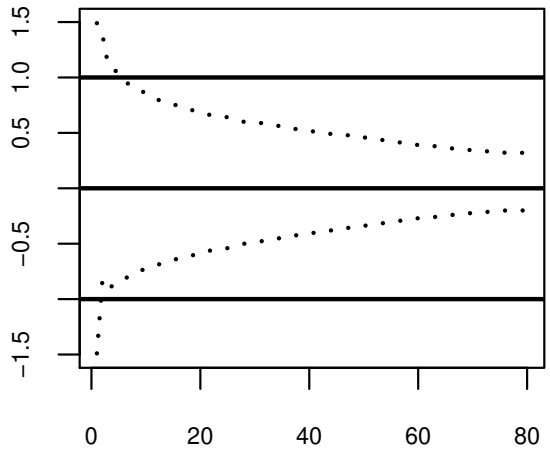


Figure 8: Numerical optima for finite T and binary signals



Plots optimal responsiveness to positive and negative signals (β_H and $-\beta_L$) for the unbiased (solid lines) and optimally biased (dotted lines) cases over $1 \leq T \leq 80$. The remaining parameters are fixed at $\mu_0 = 0.5$, $c \sim U[0,1]$, $b(\hat{\mu}) = \frac{1}{4}\hat{\mu}$, $p = 0.75$, $q = 0.25$

make it impossible to keep logit-beliefs close to μ_L^*. Although the agent's mean logit-belief in the low state stays close to μ_L^*, her realized logit-belief will typically be either very small or very large. Since $L(0) = 0$ and $L(1) < 0$ this is costly; the low-type agent would in fact be worse off than under unbiased Bayesian updating. Conservatism addresses this problem by keeping the low-type agent's beliefs close to μ_L^* in probability. The proof of Proposition 2 formalizes this intuition: it shows that any updating process that is not both conservative and downward-neutral biased must do strictly worse than a process that is, and that an optimal updating process allows the agent to closely approximate her "first best" payoffs by keeping her belief bounded away from zero at μ_L^* in the low state while still learning her type rapidly in the high state.

While Proposition 2 characterizes optimal behavior for large T, we can also characterize the finite-T case numerically. Figure 8 shows the optimal updating policy over the range $1 \leq T \leq 80$ for a binary signals example with a uniform cost distribution, an objective prior of $\mu_0 = \frac{1}{2}$ and

belief utility $b(\hat{\mu}) = \frac{1}{4}\hat{\mu}$. These parameters satisfy the long-term learning condition $L(1) < 0$ and imply $\mu_L^* = \frac{1}{4}$: the agent would like to maintain a confidence level of 25% in the low state. As in our experiment, signals are accurate with probability 0.75. The agent is optimally asymmetric over the entire range, conservative for $T > 8$, and increasingly conservative as T increases.

6.3 Robustness of Biased Bayesian Updating

The optimal updating rule β^T that we characterized in Proposition 2 depends on the specific decision problem (summarized by per-period utilities $L(\hat{\mu})$ and $H(\hat{\mu})$). However, we can show that this dependence is weak in the following sense: if the agent faces a new decision problem $(\tilde{L}, \tilde{H})$ and continues to use the old updating rule β^T, then she can do almost as well as when she uses the new optimal updating rule $\tilde{\beta}^T$.

Proposition 3. *Fix a signal distribution (F_H, F_L). Consider two decision problems (L, H) and $(\tilde{L}, \tilde{H})$ with optimal updating rules β^T and $\tilde{\beta}^T$, respectively. Assume that the agent uses the updating rule β^T for the latter problem. Then the agent's combined utility and subjective belief at any relative time τ converge in probability to the first-best values as $T \to \infty$.*

The result implies that the agent can do very well by applying a uniform updating bias (independent of the decision problem) and by choosing an initial subjective prior close to the low-type's optimal belief. This observation allows for the possibility of an evolutionary process in which Nature selects an updating rule for a generic decision problem which the agent then applies to different specific problems throughout life.

7 Conclusion

We use a large-scale experiment to characterize belief updating in a setting where ego is at stake. We document two biases: participants are asymmetric as they respond more to positive than negative information, and conservative by overall responding less to feedback than a Bayesian. It seems plausible that asymmetric updating is a bias rather than a cognitive error. This is, however, less obvious for conservatism. A control treatment where participants update on a random event rather than an ego relevant one provides evidence that conservative updating is also a behavioral bias. The second experiment shows that beliefs have a causal impact on economic choices. More importantly, we show that more and less conservative participants respond equally to an equivalent change in their confidence. If biased beliefs were only a psychological phenomenon without economic consequences, then more conservative updating led to a *greater* response to beliefs. Another piece of evidence for the economic consequences of biased updating is that more conservative participants have less self-knowledge about their ability than others. Finally, we provide a simple model illustrating how asymmetric and conservative updating can be complementary techniques for self-esteem management.

Our results on updating biases and their link to decision-making point towards several practical applications. First of all, policy makers should take updating biases into account when communicating feedback. Two sequences of signals that have the same Bayesian information content (such as many weak signals versus one strong signal) might induce very different belief changes. This issue is salient in labor markets, for example, given the recent shift towards more frequent performance reviews (Church et al., 2012). Feedback may also need to be tailored to the worker if (as subsequent work suggests) there is variation in how they process it. We have already seen (above) that more conservative types tend to have less informative beliefs, and are thus less likely to be either over- or under-confident even given the *same* sequence of signals. Correlating updating types with relevant real-world behaviors such as overconfidence, risk-taking and competitiveness might help explain some of these disparate phenomena as manifestations of updating biases.

Gender differences is an important special case of heterogeneity. We briefly summarize gender differences in our data in the supplemental materials (S-1). We see that men are more confident than women both unconditionally and conditional on their true ability, significantly less conservative when updating, and less averse to feedback. Interpreted through the lens of our model, these patterns are consistent with women placing a relatively high value on belief utility. Together the results and the model may thus help to explain why male participants tend to report higher self-confidence than women (Barber and Odean, 2001) and show a greater willingness to enter competitions (Niederle and Vesterlund, 2007).

While we focus on belief updating here due to space constraints, we have also found suggestive evidence that selective acquisition of information plays a role in self-confidence management. In the supplemental materials (S-2) we show that a sizeable minority of our participants are strictly averse to feedback, that low confidence has a causal effect on aversion, and that this is consistent with the theoretical framework in Section 6. This provides support for models that emphasize selective acquisition, such as Kőszegi (2006), and complements work documenting cognitive errors in information acquisition decisions such as Descamps et al. (2021). It also opens the broader question as to what kind of feedback or information sources agents would choose. For example, is there a link between how much agents bias their beliefs and the extent to which they demand information from sources that differ by how biased an information they provide.

References

Adebambo, Biljana N. and Xuemin (Sterling) Yan, "Momentum, Reversals, and Fund Manager Overconfidence," *Financial Management*, 2016, *45* (3), 609–639.

Akerlof, George A. and William T. Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, 1982, *72 (3)*, 307–319.

Allen, Franklin, "Discovering personal probabilities when utility functions are unknown," *Management Science*, 1987, *33* (4), 542–544.

Arellano, Manuel and Bo Honore, "Panel data models: some recent developments," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, Vol. 5 of *Handbook of Econometrics*, Elsevier, 2001, chapter 53, pp. 3229–3296.

_ **and Stephen Bond**, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, April 1991, *58* (2), 277–97.

Barber, Brad M. and Terrance Odean, "Boys Will Be Boys: Gender, Overconfidence, And Common Stock Investment," *The Quarterly Journal of Economics*, February 2001, *116* (1), 261–292.

Barron, Kai, "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?," *Experimental Economics*, Mar 2021, *24* (1), 31–58.

Benabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 2002, *117 (3)*, 871–915.

Benjamin, Daniel J., "Chapter 2 - Errors in probabilistic reasoning and judgment biases," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics - Foundations and Applications 2*, Vol. 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, North-Holland, 2019, pp. 69–186.

Benoit, Jeanâ Pierre and Juan Dubra, "Apparent Overconfidence," *Econometrica*, 09 2011, *79* (5), 1591–1625.

Brocas, Isabelle and Juan D. Carrillo, "The value of information when preferences are dynamically inconsistent," *European Economic Review*, 2000, *44*, 1104–1115.

Brunnermeier, Markus K. and Jonathan A. Parker, "Optimal Expectations," *American Economic Review*, September 2005, *95* (4), 1092–1118.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Götte, and Aldo Rustichini, "Overconfidence and Social Signalling," *Review of Economic Studies*, 2013, *80* (3), 949–983.

Buser, Thomas, Leonie Gerhards, and Joël van der Weele, "Responsiveness to feedback as a personal trait," *Journal of Risk and Uncertainty*, Apr 2018, *56* (2), 165–192.

Camerer, Colin and Dan Lovallo, "Overconfidence and Excess Entry: An Experimental Approach," *The American Economic Review*, 1999, *89* (1), pp. 306–318.

Caplin, Andrew and John Leahy, "Psychological Expected Utility Theory And Anticipatory Feelings," *The Quarterly Journal of Economics*, February 2001, *116* (1), 55–79.

Carrillo, Juan D. and Thomas Mariotti, "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies*, 2000, *67*, 529–544.

Charness, Gary, Aldo Rustichini, and Jeroen van de Ven, "Self-confidence and strategic behavior," *Experimental Economics*, Mar 2018, *21* (1), 72–98.

_ **and Dan Levin**, "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect," *American Economic Review*, September 2005, *95* (4), 1300–1309.

Church, Emily, Sophie Lambin, and Larry Yu, "Delivering results: Growth and value in a volatile world," *15th Annual Global CEO Survey 2012*, 2012.

Compte, Olivier and Andrew Postlewaite, "Confidence-Enhanced Performance," *American Economic Review*, December 2004, *94* (5), 1536–1557.

Coutts, Alexander, "Good news and bad news are still news: experimental evidence on belief updating," *Experimental Economics*, Jun 2019, *22* (2), 369–395.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, "Investor Psychology and Security Market Under- and Overreactions," *Journal of Finance*, 1998, *53 (6)*, 1839–1885.

Danz, David, Lise Vesterlund, and Alistair J Wilson, "Belief Elicitation: Limiting Truth Telling with Information on Incentives," Working Paper 27327, National Bureau of Economic Research June 2020.

Descamps, Ambroise, Sebastien Massoni, and Lionel Page, "Learning to hesitate," *Experimental Economics*, June 2021.

Drobner, Christoph, "Motivated Beliefs and Anticipation of Uncertainty Resolution," *American Economic Review: Insights*, forthcoming.

Eil, David and Justin M. Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–38.

El-Gamal, Mahmoud and Daniel Grether, "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of the American Statistical Association*, 1995, *90* (432), 1137–1145.

Eliaz, Kfir and Andrew Schotter, "Paying for Confidence: an Experimental Study of the Demand for Non-Instrumental Information," *Games and Economic Behavior*, November 2010, *70* (2), 304–324.

Englmaier, Florian, "A Brief Survey on Overconfidence," in D. Satish, ed., *Behavioral Finance – an Introduction*, ICFAI University Press, 2006.

Ertac, Seda, "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback," *Journal of Economic Behavior & Organization*, 2011, *80* (3), 532–545.

Fischhoff, Baruch and Ruth Beyth-Marom, "Hypothesis Evaluation from a Bayesian Perspective," *Psychological Review*, 1983, *90 (3)*, 239–260.

Galasso, Alberto and Timothy S. Simcoe, "CEO Overconfidence and Innovation," *Management Science*, 2011, *57* (8), 1469–1484.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, November 2010, pp. 1399–1433.

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, April 1989, *18* (2), 141–153.

Gotthard-Real, Alexander, "Desirability and information processing: An experimental study," *Economics Letters*, 2017, *152*, 96–99.

Grether, David M, "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, November 1980, *95* (3), 537–57.

Grether, David M., "Testing bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior & Organization*, January 1992, *17* (1), 31–57.

Grossman, Zachary and David Owens, "An unlucky feeling: Overconfidence and noisy feedback," *Journal of Economic Behavior & Organization*, 2012, *84* (2), 510–524.

Healy, Paul J., "Explaining the BDM - or any Random Binary Choice Elicitation Mechanism - to Subjects," Working Paper 2018.

_ **and John Kagel**, "Testing Elicitation Mechanisms via Team Chat," Working Paper 2021.

Hoelzl, Erik and Aldo Rustichini, "Overconfident: Do You Put Your Money On It?," *Economic Journal*, 04 2005, *115* (503), 305–318.

Hollard, Guillaume, Sébastien Massoni, and Jean-Christophe Vergnaud, "In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments," *Theory and Decision*, Mar 2016, *80* (3), 363–387.

Holt, Charles A., "Preference Reversals and the Independence Axiom," *The American Economic Review*, 1986, *76* (3), 508–515.

_ **and Angela M. Smith**, "Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes," *American Economic Journal: Microeconomics*, 2016, *8* (1), 110–139.

Hossain, Tanjim and Ryo Okui, "The Binarized Scoring Rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

Kahneman, Daniel and Amos Tversky, "On the Psychology of Prediction," *Psychological Review*, 1973, *80* (4), 237–251.

Karni, Edi, "A Mechanism for Eliciting Probabilities," *Econometrica*, 03 2009, *77* (2), 603–606.

Kőszegi, Botond, "Ego Utility, Overconfidence, and Task Choice," *Joural of the European Economic Association*, 2006, *4* (4), 673–707.

Kogan, Shimon, Florian H. Schneider, and Roberto A. Weber, "Self-serving biases in beliefs about collective outcomes," ECON - Working Papers 379, Department of Economics - University of Zurich March 2021.

Li, Meng, Nicholas C. Petruzzi, and Jun Zhang, "Overconfident Competing Newsvendors," *Management Science*, 2017, *63* (8), 2637–2646.

Long, J Bradford De, Andrei Shleifer, and Robert Waldmann, "The Survival of Noise Traders in Financial Markets," *The Journal of Business*, January 1991, *64* (1), 1–19.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *The Journal of Finance*, 2005, *60* (6), 2661–2700.

Massey, Cade and George Wu, "Detecting Regime Shifts: the Causes of Under- and Overreaction," *Management Science*, 2005, *51* (6), 932–947.

Mayraz, Guy, "Priors and Desires—a Bayesian Model of Wishful Thinking and Cognitive Dissonance," Technical Report, University of Sydney 2019.

Miller, Dale and Michael Ross, "Self-Serving Biases in the Attribution of Causality: Fact or Fiction?," *Psychology Bulletin*, 1975, *82* (2), 213–225.

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing Self-Confidence: Theory and Experimental Evidence," Working Paper 17014, National Bureau of Economic Research May 2011.

Moore, Don A. and Paul J. Healy, "The Trouble With Overconfidence," *Psychological Review*, April 2008, *115* (2), 502–517.

Mullainathan, Sendhil, "A Memory-Based Model Of Bounded Rationality," *The Quarterly Journal of Economics*, August 2002, *117* (3), 735–774.

Nickell, Stephen J, "Biases in Dynamic Models with Fixed Effects," *Econometrica*, November 1981, *49* (6), 1417–1426.

Niederle, Muriel and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete Too Much?," *The Quarterly Journal of Economics*, August 2007, *122* (3), 1067–1101.

Odean, Terrance, "Are investors reluctant to realize their losses?," *The Journal of Finance*, 1998, *53* (5), 1775–1798.

Offerman, Theo, Joep Sonnemans, Gijs Van de Kuilen, and Peter Wakker, "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *The Review of Economic Studies*, October 2009, *76* (29), 1461–1489.

Oprea, Ryan and Sevgi Yuksel, "Social Exchange of Motivated Beliefs," *Journal of the European Economic Association*, 09 2021.

Rabin, Matthew, "Psychology and Economics," *Journal of Economic Literature*, March 1998, *36* (1), 11–46.

_ , "Inference By Believers In The Law Of Small Numbers," *The Quarterly Journal of Economics*,

August 2002, *117* (3), 775–816.

_ **and Joel Schrag**, "First Impressions Matter: A Model Of Confirmatory Bias," *The Quarterly Journal of Economics*, February 1999, *114* (1), 37–82.

Santos-Pinto, Luis and Joel Sobel, "A Model of Positive Self-Image in Subjective Assessments," *American Economic Review*, December 2005, *95* (5), 1386–1402.

Schwardmann, Peter and Joël van der Weele, "Deception and self-deception," *Nature Human Behaviour*, Oct 2019, *3* (10), 1055–1061.

Slovic, Paul and Sarah Lichtenstein, "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," *Organizational Behavior and Human Performance*, 1971, *6*, 649–744.

Stein, Charles, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1972, pp. 583–602.

Stock, James H. and Motohiro Yogo, "Testing for Weak Instruments in Linear IV Regression," in Donald W. K. Andrews and James Stock, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, 2005, p. 80–108.

Svenson, Ola, "Are We All Less Risky and More Skillful Than Our Fellow Drivers?," *Acta Psychologica*, 1981, *47*, 143–148.

Van den Steen, Eric, "Rational Overoptimism (and Other Biases)," *American Economic Review*, September 2004, *94* (4), 1141–1151.

_ , "Overconfidence by Bayesian-Rational Agents," *Management Science*, 2011, *57* (5), 884–896.

Wetzel, Christopher, "Self-Serving Biases in Attribution: a Bayesian Analysis," *Journal of Personality and Social Psychology*, 1982, *43* (2), 197–209.

Wilson, Alistair J. and Emanuel Vespa, "Paired-Uniform Scoring: Implementing a Binarized Scoring Rule with Non-mathematical Language," Working Paper 2017.

Wilson, Andrea, "Bounded Memory and Biases in Information Processing," NajEcon Working Paper Reviews, www.najecon.org April 2003.

Wiswall, Matthew and Basit Zafar, "Determinants of College Major Choice: Identification using an Information Experiment," *The Review of Economic Studies*, 12 2014, *82* (2), 791–824.

Wolosin, Robert J., Steven Sherman, and Amnon Till, "Effects of Cooperation and Competition on Responsibility Attribution After Success and Failure," *Journal of Experimental Social Psychology*, 1973, *9*, 220–235.

Zábojník, Ján, "A model of rational bias in self-assessments," *Economic Theory*, January 2004, *23* (2), 259–282.

A Proofs

A.1 Proof of Proposition 1

When $b(\hat{\mu}) = 0$ for all $\hat{\mu}$, the objective function in (9) is maximized if and only if for any possible history of signals at any time $t \leq T$ and associated Bayesian belief μ_t the following holds: $\hat{\mu}^t > c$ iff $\mu^t > c$. Since the cost distribution is continuous and positive, this implies $\hat{\mu}^t = \mu^t$ for any signal history that generates the objective Bayesian posterior μ^t. Because all signal realizations are informative (and hence occur with positive probability) we obtain for $t = 1$ already K linear equations of the form $\text{logit}(\hat{\mu}^0) + \beta_k \lambda_k = \text{logit}(\mu^0) + \lambda_k$, one for each signal realization. As we have $K+1$ unknowns we can use any of the signal realizations at time $t = 2$ – e.g. two consecutive $k = 1$ realizations – to uniquely pin down $\beta_k = 1$ and $\hat{\mu}^0 = \mu^0$.

A.2 Auxiliary Approximation Lemma

For our proofs, we will frequently exploit that logit beliefs in our model are sums of independent random variables. While these variables are i.i.d. their distribution generally depends on T (because the responsiveness function changes with T), so we cannot use the standard central limit theorem. Instead we use Stein's 1972 method to bound the approximation error of the central limit theorem in our framework.

Consider the random variable Y defined over the realizations k of a single signal:

$$Y(k) = \hat{\beta}_k \lambda_k \text{ with probability } F_L(k) \tag{13}$$

where $\hat{\beta}_k \leq 1$ is the normalized responsiveness (which implies that for at least one realization we have $\hat{\beta}_k = 1$). The following lemma will be useful:

Lemma 1. *Consider any normalized responsiveness function. Let $k^* = \arg\min_k |\lambda_k|$. We then have $Var(Y) \geq F_L(k^*)\left(1 - F_L(k^*)\right)\lambda_{k^*}$.*

Proof: The variance of Y is minimized over all normalized responsiveness functions if $\beta_{k^*} = 1$ and $\beta_k = 0$ for all $k \neq k^*$. This reduces Y to a simple Bernoulli random variable and the result follows.

We define two new constants:

$$\begin{aligned} M_L &= 5\left(\frac{\max_k \lambda_k}{\sqrt{F_L(k^*)\left(1 - F_L(k^*)\right)\lambda_{k^*}}}\right)^3 \\ M_H &= 5\left(\frac{\max_k \lambda_k}{\sqrt{F_H(k^*)\left(1 - F_H(k^*)\right)\lambda_{k^*}}}\right)^3 \end{aligned}$$

We can now prove the following approximation for subjectve beliefs:

Lemma 2. *Let $\epsilon > 0$ and $-\infty \leq a < b \leq \infty$. The random variable $W = \frac{logit(\hat{\mu}_{\lfloor \tau T \rfloor}) - \hat{\gamma}^L_{\lfloor \tau T \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T \rfloor}}$ satisfies:*

$$Prob(a \leq W \leq b | L) \leq \Phi(b + 2\epsilon) - \Phi(a - 2\epsilon) + \frac{M_L}{\epsilon\sqrt{\tau T}}$$

where Φ is the cdf of the normal distribution $N(0,1)$. An analogous result holds for beliefs in the high state where M_L is replaced by M_H.

Note, that the upper bound depends only on ϵ, τT and the distribution of the signal distribution but (importantly) *not* on the particular responsiveness function.

Proof: WLOG we focus on low-state beliefs only. We define the function h:[39]

$$h(x) = \begin{cases} 0 & \text{if } x < a - 2\epsilon \\ \frac{1}{2\epsilon^2}(x - a + 2\epsilon)^2 & \text{if } a - 2\epsilon \leq x < a - \epsilon \\ 1 - \frac{1}{2\epsilon^2}(x - a)^2 & \text{if } a - \epsilon \leq x < b \\ 1 & \text{if } a \leq x < b \\ 1 - \frac{1}{2\epsilon^2}(x - b)^2 & \text{if } b \leq x < b + \epsilon \\ \frac{1}{2\epsilon^2}(x - b - 2\epsilon)^2 & \text{if } b + \epsilon \leq x < b + 2\epsilon \\ 0 & \text{if } b + 2\epsilon \leq x \end{cases}$$

This function approximates the indicator function that takes value 1 on the interval $[a, b]$ such that h is bounded above by the indicator function on the interval $[a - 2\epsilon, b + 2\epsilon]$, bounded below by the indicator function on $[a, b]$ and bounded derivative $|h'(x)| \leq \frac{1}{\epsilon}$. Now we use Stein's inequality to establish

$$|\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]| \leq \frac{\max_x h'(x) 5E|X_i|^3}{\sqrt{\tau T}}$$

where $Z \sim N(0, 1)$ and X_i are i.i.d. random variables of the form $X = \frac{Y - E(Y)}{\sqrt{Var(Y)}}$. Thus

$$Prob(a \leq W \leq b) \leq \mathbb{E}[h(W)] \leq \mathbb{E}[h(Z)] + \frac{(\max_x h'(x))\, 5E|X_i|^3}{\sqrt{\tau T}}$$

and the result of the lemma then follows.

A.3 Uniform Downward-Neutral Bias

We define a particular responsiveness function which we call the *uniform downward neutral bias* that approximates the utility of the unrestricted agent who can freely choose her beliefs in both states of the world. This will be useful to prove proposition 2 where we show that non-conservative responsiveness functions or those which do not satisfy the DNB property cannot be optimal because they cannot approximate the utility of the unrestricted agent.

For a given signal distribution, we partition the set of possible realizations into an "Up-set" $U = \{k|\lambda_k > 0\}$ and a "Down-set" $D = \{k|\lambda_k < 0\}$. We fix a constant $\frac{1}{2} < \theta < 1$. For each T we define the following biased Bayesian updating process:

$$\begin{aligned} \hat{\mu}_0^T &= \mu_L^* \\ \beta_k &= \begin{cases} T^{-\theta} & \text{for } k \in U \\ T^{-\theta} \underbrace{\frac{\sum_{k \in U} F_L(k)\lambda_k}{-\sum_{k \in D} F_L(k)\lambda_k}}_{\kappa} & \text{for } k \in D \end{cases} \end{aligned} \tag{14}$$

Note, that $0 < \kappa < 1$ because the unbiased agent's expected change in logit-beliefs in the low state has to be negative (hence, $\sum_{k \in U} F_L(k)\lambda_k + \sum_{k \in D} F_L(k)\lambda_k < 0$). We can derive the mean and

[39] For $a = -\infty$ ($b = \infty$) we adapt the definition naturally and let $h(x) = 1$ for $x < b$ ($x > a$).

variance of logit-beliefs at relative time τ in both states:

$$\begin{aligned}
\hat{\gamma}^H_{\tau T} &= \text{logit}(\mu^*_L) + \tau T^{1-\theta} \underbrace{\left(\sum_{k \in U} F_H(k)\lambda_k + \kappa \sum_{k \in D} F_H(k)\lambda_k \right)}_{\Gamma_H} \\
\hat{\gamma}^L_{\tau T} &= \text{logit}(\mu^*_L) \\
\left(\hat{\sigma}^H_{\tau T}\right)^2 &= \tau T^{1-2\theta} \underbrace{\left(\sum_{k \in U} F_H(k)\lambda_k^2 + \kappa^2 \sum_{k \in D} F_H(k)\lambda_k^2 - \Gamma_H^2 \right)}_{\Sigma_H > 0} \\
\left(\hat{\sigma}^L_{\tau T}\right)^2 &= \tau T^{1-2\theta} \underbrace{\left(\sum_{k \in U} F_L(k)\lambda_k^2 + \kappa^2 \sum_{k \in D} F_L(k)\lambda_k^2 \right)}_{\Sigma_L > 0} \qquad (15)
\end{aligned}$$

Note, that $\Gamma_H > 0$ because the unbiased agent's expected change in logit-beliefs in the high state is strictly positive (hence, $\sum_{k \in U} F_H(k)\lambda_k + \sum_{k \in D} F_H(k)\lambda_k > 0$) and $\kappa < 1$. We call this particular updating process the uniform downward-neutral bias (uniform DNB) because a uniform bias factor is applied to up and down signal realizations, respectively, and logit-beliefs for the low type follow a random walk without drift.

Lemma 3. *Assume a biased Bayesian with uniform DNB. At any relative time $\tau > 0$, the agent's high state belief converges in probability to* 1 *while the agent's low state belief converges in probability to μ^*_L. The total utility (9) of the agent converges to the total utility of an unrestricted agent with belief μ^*_L in the low state and belief* 1 *in the high state.*

Figure 7 illustrates the intuition for the lemma. In the high state, the agent's logit-belief at relative time τ is of order $\tau T^{1-\theta}$ according to (15). This expression converges to infinity. In the low state, the agent's logit-belief behaves like a driftless random walk whose standard deviation is of order $\sqrt{\tau} T^{\frac{1}{2}-\theta}$, which converges to 0.

To formalize this argument, we first show that for any lower bound m the probability that the high type's logit-belief lies above m at relative time τ converges to 1 as $T \to \infty$:

$$\begin{aligned}
P(\text{logit}(\hat{\mu}_{\lfloor \tau T \rfloor}) < m | H) &= P\left(\frac{\text{logit}(\mu_{\lfloor \tau T \rfloor}) - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\hat{\sigma}^H_{\tau T}} < \frac{m - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_H}} | H \right) \\
&\leq \Phi\left(\frac{m - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_H}} + 2\epsilon \right) + \frac{M_H}{\epsilon \sqrt{\tau T}}
\end{aligned}$$

For the last inequality we use our approximation lemma 2 with $a = -\infty$ and any $\epsilon > 0$. We now exploit the fact that $\frac{m - \hat{\gamma}^H_{\lfloor \tau T \rfloor}}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_H}} \to -\infty$, which holds since $\hat{\gamma}^H_{\lfloor \tau T \rfloor} \to \infty$ and the numerator is of order $O(\tau T^{1-\theta})$ while the denominator is only of order $O(\sqrt{\tau} T^{\frac{1}{2}-\theta})$.

We next show that for any $\epsilon' > 0$ the probability that the low type's belief stays within an ϵ'-neighborhood around $\text{logit}(\mu^*_L)$ converges to 1 in probability as $T \to \infty$. Note, that the expected

logit-belief at any relative time τ is $\text{logit}(\mu_L^*)$ under the uniform DNB:

$$\begin{aligned}
& P(|\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu_L^*)| > \epsilon' | L) = \\
= \; & P\left(\frac{\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} < -\frac{\epsilon'}{\hat{\sigma}_{\tau T}^L} | L \right) + P\left(\frac{\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} > \frac{\epsilon'}{\hat{\sigma}_{\tau T}^L} | L \right) \\
\leq \; & \Phi\left(\frac{-\epsilon'}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_L}} + 2\epsilon \right) + 1 - \Phi\left(\frac{\epsilon'}{\sqrt{\tau} T^{\frac{1}{2}-\theta} \sqrt{\Sigma_L}} - 2\epsilon \right) + \frac{2M_L}{\epsilon \sqrt{\tau T}}
\end{aligned}$$

For the last inequality we fix any $\epsilon > 0$ and use our approximation lemma 2 twice. We can make this upper bound as small as we want for sufficiently high T since $\theta > \frac{1}{2}$.

Also note that we can obtain a uniform upper bound for all relative time by setting $\tau = 1$ on the RHS. Since the cost distribution is atomless, it follows that the expected utility of the low type agent converges to the utility of the unconstrained low type with constant belief μ_L^*.

A.4 Proof of Proposition 2

Step 1: Conservatism We first show conservatism (claim 1 of the proposition) through proof by contradiction. The intuition for conservatism is as follows: assume the agent's responsiveness does not converge to 0. There will be some realization k and a sequence (T^j), such that $|\beta_k^{T^j}| > \delta > 0$ for some $\delta > 0$. We will show that the agent's total utility in the low state converges to at most 0 as $T^j \to \infty$. According to lemma 3 an agent with uniform DNB would do strictly better: hence the agent cannot be optimally biased.

We start by bounding the probability that subjective beliefs fall within the interval $[\epsilon', 1-\epsilon']$ in the low state:

$$\begin{aligned}
& P(\epsilon' < \hat{\mu}_{\lfloor \tau T^j \rfloor} < 1 - \epsilon' | L) \\
= \; & P\left(\frac{\text{logit}(\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} < \frac{\text{logit}(\hat{\mu}^{\lfloor \tau T \rfloor}) - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} < \frac{\text{logit}(1-\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} | L \right) \\
\leq \; & \Phi\left(\frac{\text{logit}(1-\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} + 2\epsilon \right) - \Phi\left(\frac{\text{logit}(\epsilon') - \text{logit}(\mu_L^*)}{\hat{\sigma}_{\tau T}^L} - 2\epsilon \right) + \frac{M_L}{\epsilon \sqrt{\tau T}}
\end{aligned}$$

For the last inequality we fix any $\epsilon > 0$ and use our approximation lemma 2. We next replicate the proof of lemma 1 to show:

$$\hat{\sigma}_{\tau T^j}^L \geq \sqrt{\tau T^j} \underbrace{\sqrt{F_L(k^*)\left(1 - F_L(k^*)\right) \lambda_{k^*} \delta}}_{M' > 0}$$

We can therefore simplify the upper bound:

$$P(\epsilon' < \hat{\mu}_{\lfloor \tau T^j \rfloor} < 1 - \epsilon' | L) \leq \frac{1}{\sqrt{2\pi}} \left(\frac{\text{logit}(1-\epsilon') - \text{logit}(\epsilon')}{\sqrt{\tau T^j} M'} + 4\epsilon \right) + \frac{M_L}{\epsilon \sqrt{\tau T^j}} = M'' \epsilon + \frac{M'''(\epsilon, \epsilon')}{\sqrt{\tau T^j}}$$

Now fix a relative time τ^*. We can bound the total utility of the low type above by $\tau^* b(1) + (1 - \tau^*) K$

where

$$K = \underbrace{\left(M''\epsilon + \frac{M'''(\epsilon,\epsilon')}{\sqrt{\tau T^j}}\right) b(1)}_{\text{Bound on expected utility from posterior falling within } [\epsilon', 1-\epsilon'] \text{ after relative time } \tau^*} + \underbrace{b(\epsilon')}_{\text{Bound on expected utility from posteriors below } \epsilon' \text{ after relative time } \tau^*} + \underbrace{A\left[b(1) - \int_0^{1-\epsilon'} c dG(c)\right]}_{\text{Bound on expected utility from posteriors above } 1-\epsilon' \text{ after relative time } \tau^* \text{ (probability A)}}$$

Due to the fact that the cost distribution is non-atomic, the last term is negative for sufficiently small ϵ' as $L(1) < 0$. Next, choose first τ^* and ϵ' and then T^* to make $\tau^* b(1)$ and the first two terms of K as small as desired for all $T^j > T^*$. Therefore, the low type's utility cannot be bounded away from 0 and the biased Bayesian does not do strictly better than an unbiased Bayesian for large T^j.

Step 2: DNB The proof of claim 2 of the proposition proceeds in 2 sub-steps. (A) We first show that for any constant $M > 0$ we have $\max_k \beta_k^T > \frac{M}{T}$ for any sufficiently large T. (B) Next, if optimal updating does not exhibit DNB for large T then the mean logit low-type belief converges either to plus or minus infinity. In both cases, the biased agent's utility will be strictly lower than under the uniform DNB.

We start with part A. Assume this claim is wrong. Then, we can find some M and a sub-sequence T^j such that $\max_k \beta_k^{T^j} < \frac{M}{T^j}$. This implies that mean logit-belief in the high state at any relative time τ is bounded above by $M^* = M \max_k \lambda_k$. But since belief utility is strictly increasing, her utility will be strictly lower than the utility of the unrestricted agent, and therefore also strictly lower than for the agent with uniform DNB for any large enough T. This is a contradiction since we assumed that the responsiveness function is optimal.

Next consider claim B. Assume that $\sum_k F_L(k)\hat{\beta}_k^T \lambda_k$ does not converge to 0. Then there is some $\epsilon > 0$ and a sub-sequence T^j such that $|\sum_k F_L(k)\hat{\beta}_k^{T^j}\lambda_k| > \epsilon$. For any constant M, this implies $|\sum_k F_L(k)\beta_k^{T^j}\lambda_k| > \frac{M\epsilon}{T^j}$ as long as T^j is sufficiently big. Hence, the mean logit-belief of the low type converges either to $-\infty$ or $+\infty$.

We fix $\tau^* < 1$ and look at the case $\hat{\gamma}^L_{\lfloor \tau^* T^j \rfloor} \to -\infty$ first. Take a constant $B < \text{logit}(\mu_L^*)$. We use our approximation lemma 2 (for some $\epsilon > 0$):

$$\begin{aligned}
P(\text{logit}(\hat{\mu}_{\lfloor \tau^* T^j \rfloor}) > B | L) &= P\left(\frac{\text{logit}(\hat{\mu}_{\lfloor \tau^* T^j \rfloor}) - \hat{\gamma}^L_{\lfloor \tau^* T^j \rfloor}}{\hat{\sigma}^L_{\tau^* T^j}} > \frac{B - \hat{\gamma}^L_{\lfloor \tau^* T^j \rfloor}}{\hat{\sigma}^L_{\tau^* T^j}} | L\right) \\
&\leq 1 - \Phi\left(\frac{B - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\tau^* T^j}} - 2\epsilon\right) + \frac{M_L}{\epsilon\sqrt{\tau^* T^j}} \\
&\leq 1 - \Phi(-2\epsilon) + \frac{M_L}{\epsilon\sqrt{\tau^* T^j}} \\
&\leq \frac{2}{3} \quad \text{for } \epsilon \text{ small enough and large enough } T^j
\end{aligned}$$

Hence, the probability of the low-type's logit-belief being below B for relative times $\tau > \tau^*$ is at least $\frac{1}{3}$. Hence, the low-type's utility is strictly lower than for an agent with unrestricted beliefs. This is a contradiction since we assumed that the responsiveness function is optimal. We can arrive at a similar contradiction for the case $\hat{\gamma}^L_{\lfloor \tau^* T^j \rfloor} \to \infty$.

Step 3: Initial Beliefs

We prove claims 3 and 4 of proposition 2 in 3 sub-steps. (A) We define an upper envelope function $U(x)$ for $L(x)$. (B) We show that $\hat{\sigma}^L_{\tau T} \to 0$ as $T \to \infty$, which is a strong form of conservatism. (C) We show that this implies claims (3) and (4) of proposition 2.

We start with part A. Using Taylor's theorem we can write

$$L(x) = L(\mu^*_L) + \frac{1}{2}L''(y)(x - \mu^*_L)^2 \tag{16}$$

for some $y \in [x, \mu^*_L]$. Note that L'' is continuous and hence strictly negative in an ϵ-neighborhood of μ^*_L, since $L''(\mu^*_L) < 0$. We can assume that $L''(y) \leq -A$ for some $A > 0$ in that neighborhood. We can now define the upper envelope function $U(x)$ for $L(x)$ as follows:

$$U(x) = \begin{cases} L(\mu^*_L) - \frac{A}{2}(\mu^*_L - \epsilon)^2 & \text{for } x \leq \mu^*_L - \epsilon \\ L(\mu^*_L) - \frac{A}{2}(x - \mu^*_L)^2 & \text{for } \mu^*_L - \epsilon \leq x \leq \mu^*_L + \epsilon \\ L(\mu^*_L) - \frac{A}{2}(\mu^*_L + \epsilon)^2 & \text{for } x \geq \mu^*_L + \epsilon \end{cases} \tag{17}$$

This upper envelope will lie above $L(x)$ in the ϵ-neighborhood. We can refine the upper envelope function such that the upper envelope function dominates $L(x)$ on the interval $[0, 1]$ by considering the following set M that includes all local maxima outside the ϵ-neighborhood:

$$M = \big\{x | L'(x) = 0\big\} \setminus [\mu^*_L - \epsilon, \mu^*_L + \epsilon]$$

Denote the supremum of the $L(M)$ with m^*. Due to the Bolzano-Weierstrass theorem, there is a sequence $(x^j) \subset M$ such that $L(x^j)$ converges to m^*. Due to continuity, there is a subsequence $(x^{j'})$ of (x^j) and a $\tilde{x}$ such that $x^{j'} \to \tilde{x}$ and $L(x^{j'}) \to m^*$ and $L(\tilde{x}) = m^*$. If $m^* \geq L(\mu^*_L)$ then we get a contradiction because we assumed that the maximum at μ^*_L is unique. Hence, $m^* < L(\mu^*_L)$. Therefore, we can simply make the ϵ-neighborhood of the upper-envelope function small enough such that it always lies above m^*. This will ensure that the upper envelope function dominates L on the interval $[0, 1]$.[40]

For part B, assume that $\hat{\sigma}^L_T$ does not converge to 0 as $T \to \infty$. Then there is a subsequence (T^j) and some $\delta > 0$ such that $\hat{\sigma}^L_{T^j} > \delta$. Let $\delta' < \frac{\delta\sqrt{2\pi}}{4}$ and $\tau^* < 1$. We use our approximation lemma 2 (for some $\epsilon > 0$ and any $\tau > \tau^*$):

$$\begin{aligned}
& P(|\text{logit}(\hat{\mu}_{\lfloor \tau T^j \rfloor}) - \text{logit}(\mu^*_L)| < \delta' | L) \\
= \; & P\left(\frac{\text{logit}(\mu^*_L) - \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} < \frac{\text{logit}(\hat{\gamma}^L_{\lfloor \tau T^j \rfloor}) - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} < \frac{\text{logit}(\mu^*_L) + \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} | L \right) \\
\leq \; & \Phi\left(\frac{\text{logit}(\mu^*_L) + \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} + 2\epsilon \right) - \Phi\left(\frac{\text{logit}(\mu^*_L) - \delta' - \hat{\gamma}^L_{\lfloor \tau T^j \rfloor}}{\hat{\sigma}^L_{\lfloor \tau T^j \rfloor}} - 2\epsilon \right) + \frac{M_L}{\epsilon\sqrt{\tau T^j}} \\
\leq \; & \frac{1}{\sqrt{2\pi}} \left(\frac{2\delta'}{\hat{\sigma}^L_{T^j}} + 4\epsilon \right) + \frac{M_L}{\epsilon\sqrt{\tau T^j}} \\
\leq \; & \frac{1}{2} + \frac{4\epsilon}{\sqrt{2\pi}} + \frac{M_L}{\epsilon\sqrt{\tau^* T^j}} \\
\leq \; & \frac{2}{3} \quad \text{for } \epsilon \text{ small enough and large enough } T^j
\end{aligned}$$

[40] If there are finitely many local maxima, then the argument simplifies to m^* being the second-highest maximum.

Hence, the probability that subjective beliefs fall outside the interval $[\text{logit}^{-1}(\text{logit}(\mu_L^*-\delta')), \text{logit}^{-1}(\text{logit}(\mu_L^*+\delta'))]$ for $\tau > \tau^*$ is at least 1/3. The utility of the low-type agent using the upper-envelope function $U(x)$ accumulated over time $\tau > \tau^*$ is always strictly worse than the utility of the agent with a uniform DNB who can maintain beliefs arbitrarily closely to the optimal μ_L^*. Since her actual utility is even lower, we can strictly improve the agent's utility by using a uniform DNB. This is a contradiction since we assumed that the responsiveness function is optimal. Hence we proved $\hat{\sigma}_T^L \to 0$.

It follows that $\hat{\mu}_0^T \to \mu_L^*$. Otherwise, there would be a δ-neighborhood of μ_L^* and a subsequence (T^j) such that the initial prior $\hat{\mu}_0^{T^j}$ falls outside that interval. Combined with part A, this would imply that the agent's utility is strictly lower than under the uniform DNB along this sequence for large T^j which is a contradiction.

Combining part A with claim (3) of the proposition we immediately get convergence of low-type beliefs at any relative time τ to μ_L^*. Part A of step 2 also establishes that high-type mean-logit beliefs converge to $+\infty$. It is easy to see that $\hat{\sigma}_T^L \to 0$ implies $\hat{\sigma}_T^H \to 0$. Using lemma 2 then establishes that high-type beliefs converge to 1 in probability at any relative time $\tau > 0$.

A.5 Proof of Proposition 3

We have established in step 3 of the proof of proposition 2 that $\hat{\sigma}_T^L \to 0$. Using lemma 2 we can show that the probability that the low-type's beliefs remain in an interval around the new optimal low-type beliefs converges to 1 for any relative time τ. High-type belief convergence to 1 at all relative times is not affected by choosing a different prior.

Supplementary Material to: "Managing Self-Confidence: Theory and Experimental Evidence"

October 18, 2021

S-1 Gender Differences

By connecting different information-processing biases, our model provides one candidate framework for analysing heterogeneity in information-processing across individuals. Gender is a particularly relevant dimension. Gender differences related to self-confidence have been demonstrated in numerous studies in psychology, and economists have recently begun to investigate gender differences in beliefs about relative ability.[1] Consistent with prior work, men in our sample are significantly more confident than women: the mean difference in confidence prior to taking the quiz was 6.7 percentage points ($p < 0.001$). Some of this may reflect differences in actual ability, as men scored 7.9 on average while women scored 6.9 ($p < 0.001$). Even when we look within groups of participants who took the same version of the quiz and received the same score, we find that men are 5.0 percentage points more confident on average ($p < 0.001$).

Of course, the point of our design is not to generate additional (albeit clean) evidence of gender differences in confidence, but rather to examine what is at the root of this finding. Do women and men simply differ in their prior, or do they process information differently, or have different demands for information? To quantify gender differences in information processing, Table S-1 reports estimates of Equation 5 differentiated by gender and estimated using both OLS and instrumental variables. Men are substantially less conservative than women, reacting significantly more to both positive and negative feedback and 21% more to feedback on average (23% when estimated by IV). Estimated changes in relative asymmetry are less stable; OLS and IV point estimates of $\frac{\beta_H+\beta_H^{Male}}{\beta_L+\beta_L^{Male}} - \frac{\beta_H}{\beta_L}$ are 0.05 and -0.10, respectively, and neither is significantly different from zero ($p = 0.64, 0.74$). The evidence thus suggests that women are the more ego-defensive gender; they do not merely have different priors, but seem to process information differently. Moreover since

[1] Numerous psychology studies purport to show that men are more (over-)confident than women; see the references in Barber and Odean (2001), who use gender as a proxy measure of overconfidence in studying investment behavior. Niederle and Vesterlund (2007) show that men are much more competitive than women and that part of this difference is attributable to differences in self-confidence. They also speculate that gender differences in feedback aversion may have further explanatory power.

Table S-1: Heterogeneity in Updating by Gender

Regressor	OLS
β_H	0.346
	$(0.018)^{***}$
β_L	0.254
	$(0.013)^{***}$
β_H^{Male}	0.052
	$(0.027)^{*}$
β_L^{Male}	0.074
	$(0.024)^{***}$
N	2448
R^2	0.407

Each column is a separate regression. The outcome in all regressions is the log belief ratio. δ, β_H, and β_L are the estimated effects of the prior belief and log likelihood ratio for positive and negative signals, respectively. δ^{Male}, β_H^{Male}, and β_L^{Male} are the differential responses attributable to high ability. Robust standard errors clustered by individual reported in parentheses. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

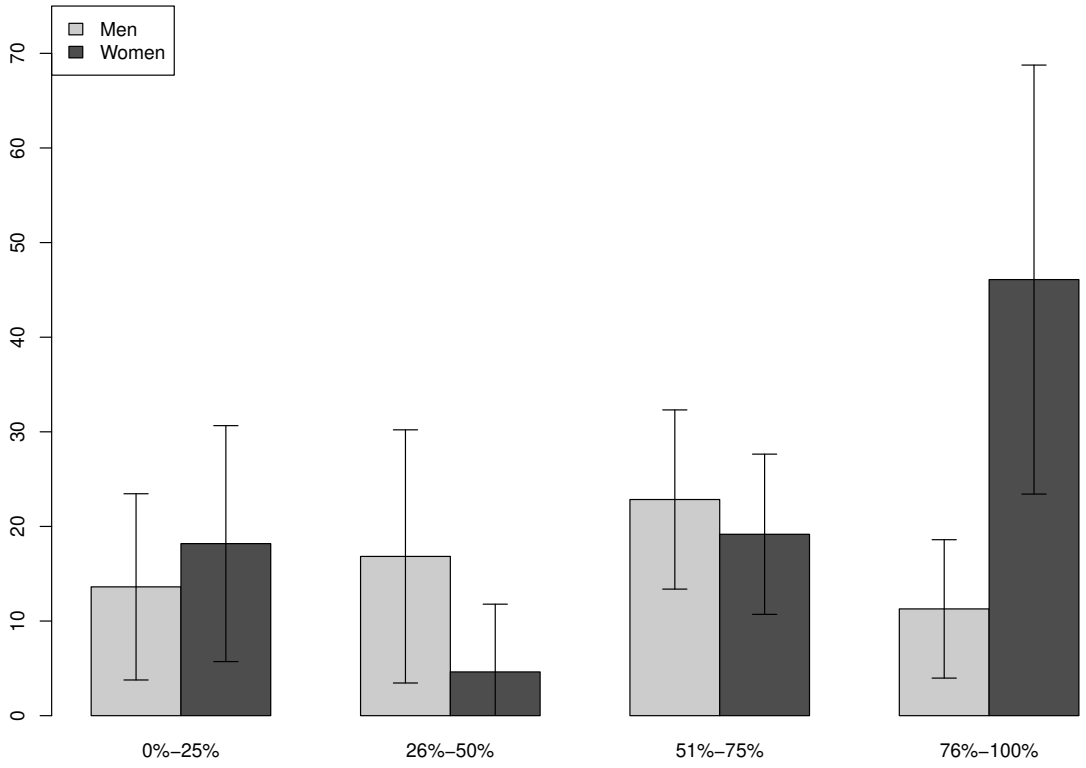
ability is uncorrelated with asymmetry and conservatism (Table 3) these gender differences cannot simply capture differences in ability.

Turning to demand for feedback, men and women place similar average valuations on information; the means reported in Table S-2 are not statistically different from each other. Men, however, are significantly less averse to feedback. They are 3.6 percentage points less likely to place negative bids for coarse information, relative to a baseline of 11% for women ($p = 0.09$). They are also 4.6 percentage points less likely to place negative bids for precise information, relative to a baseline of 11% for women ($p = 0.03$). Figure S-1 provides a less parametric view, plotting mean information values by gender and by quartile of the posterior belief distribution. The relationship between beliefs and valuations is inverse-U shaped for men, as a standard model of information demand would predict. For women, however, valuations decline somewhat from the first to second quartile and then increase dramatically from there to the fourth quartile. Confident women express significantly stronger demand for information than confident men. Interestingly, valuations are particularly low for women with beliefs between 26% and 50% (though not between 0% and 25%), similar to the pattern in Figure S-2. Overall the information demand data, like the updating data, are consistent with our theoretical framework if women are more likely than men to value belief utility.

S-2 Demand for Information

While some behavioral models do study agents who can skew their interpretation of feedback, others focus on selective *acquisition* of feedback as a technique for self-confidence management. In this section we test for this mechanism in our data and then relate it to our theoretical framework.

Figure S-1: Information Values by Beliefs and by Gender



Plots, for male and female participants separately and for quartiles of the posterior belief distribution, the mean valuations for learning whether or not the participant scored in the top half of performers.

S-3.1 Evidence

A core distinction between standard models of learning and behavioral models with belief utility (broadly defined) is that in the former agents always weakly value more information, while in the latter they may be strictly averse to it. To examine this property in our data, we calculate participants' implied value for the various information packages offered to them. For example, a participant's valuation for learning whether or not she was in the top half is defined as her bid for \$2 and learning this information minus her bid for \$2, all in cents. We take this difference to remove potential bias due to misunderstanding the dominant strategy in the "bid for \$2" decision problem.[2] Participants also bid on more precise information: learning their exact quantile. Table S-2 summarizes the results. Participants' mean value for coarse information is 16.5 (s.d. 47.8), with 9% of participants reporting a negative value. The mean valuation for precise information is higher

[2]Among our participants, 89% bid less than \$2, and 80% bid less than \$1.99.

Table S-2: Implied Valuations for Information: Summary Statistics

	N	Mean	Std. Dev.	$P(v < 0)$
Estimation Sample				
Learning top/bottom half	650	16.5	47.8	0.09
Learning percentile	650	40.0	78.3	0.09
Women				
Learning top/bottom half	338	16.4	49.8	0.11
Learning percentile	338	38.7	82.0	0.11
Men				
Learning top/bottom half	312	16.7	45.5	0.07
Learning percentile	312	41.5	74.1	0.06

Values for information are the differences between participants' bids for \$2 and their bids for the bundle of \$2 and receiving an email containing that information. Values are in cents. The final column reports the fraction of observations with strictly negative valuations. There are fewer than 656 observations because 6 participants did not provide valuations for information.

at 40.0 (s.d. 78.3), but again 9% of participants report a negative value.[3]

Result 9 (Information Aversion)**.** *A substantial fraction of participants are willing to pay to avoid learning their type.*

One caveat is that negative valuations could be an artefact of noise in participants' responses. The strongest piece of evidence that this is not the case is our next result, which shows that confidence has a causal effect on the propensity for aversion. Another clue is the high correlation ($\rho = 0.77$) between having a negative valuation for coarse information and a negative valuation for precise information, which suggests that both measures contain meaningful information. In unreported results we have developed this idea formally and shown that under the structural assumption of i.i.d. normal measurement error the bid data reject the null hypothesis of no aversion (results available on request).

Result 10. *More confident participants are causally less information-averse.*

To examine whether information aversion is more pronounced among more or less confident participants we regress an indicator $I(v_i \geq 0)$ on participants' logit posterior belief after all four rounds of updating, which is when they bid for information. Columns I–III of Table S-3 show that participants with higher posterior beliefs are indeed significantly more likely to have (weakly) positive information values. The point estimate is slightly larger and remains strongly significant when we control for ability (Column II) and gender and age (Column III). There could, however, be some other unobserved factor orthogonal to these controls that explains the positive correlation. To address this issue Columns IV and V report instrumental variables estimates. We use two instruments. First, the average score of other participants randomly assigned to the same quiz

[3]Interestingly, Eliaz and Schotter (2010) find that participants are willing to pay positive amounts for information (unrelated to ego) even when it cannot improve their decision-making.

Table S-3: Confidence and Positive Information Value

	OLS			IV	
Regressor	I	II	III	IV	V
logit(μ)	0.017	0.023	0.023	0.027	0.027
	(0.007)**	(0.009)***	(0.009)**	(0.016)*	(0.017)*
Top Half		-0.033	-0.035	-0.038	-0.042
		(0.028)	(0.028)	(0.034)	(0.034)
Male			0.029		0.027
			(0.023)		(0.023)
YOG			0.018		0.018
			(0.012)		(0.012)
First-Stage F-Statistic	-	-	-	118.48	113.19
N	609	609	609	609	609
R^2	0.007	0.010	0.016	-	-

Notes: Each column is a separate regression. Estimation via OLS is reported in Columns I–III and by IV in Columns IV–V using the instruments described in the text. The outcome variable in all regressions is an indicator equal to 1 if the participant's valuation for information was positive; the mean of this variable is 0.91. "Top Half" is an indicator equal to one if the participant scored above the median on his/her quiz type; "YOG" is the participant's year of graduation. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

type remains a valid instrument for beliefs, as in Section 4 above. In addition, once we control for whether or not the participant scored in the top half the number of positive signals she received during the updating stage is a valid instrument since signals were random conditional on ability. Estimates using these instruments are similar to the OLS estimates, slightly larger, and though less precise, still significant at the 10% level.

S-3.2 Theory

The result that low-confidence agents are (causally) likely to be information-averse is broadly consistent with a number of behavioral models which generate information aversion. In this section we examine more specifically how our data compare to the predictions of the model in Section 6 of the main paper. Towards this aim, extend the model and suppose that with probability $\epsilon > 0$ the agent is presented with the opportunity to purchase a perfectly informative signal at time $\tilde{T}$ just before learning the cost c for making costly investment. It is easy to calculate the unbiased Bayesian's willingness to pay for information, $WTP^{PB}(\mu_{\tilde{T}})$:

$$WTP^{PB}(\mu_{\tilde{T}}) = \mu_{\tilde{T}} \left(1 - \int_0^1 c dG(c)\right) - \int_0^{\mu_{\tilde{T}}} (\mu_{\tilde{T}} - c) dG(c) \tag{18}$$

Importantly, an unbiased Bayesian's value of information is always positive and single-peaked: the value of information is zero when the agent is very sure about her type and largest when she is the least sure. This valuation is generally suboptimal for an agent with belief utility, however, who

wishes to balance this motive against the needs of decision-making. If a low type were to learn the truth at time $\tilde{T}$ her carefully calibrated self-belief management would break down and she would enjoy no belief utility between periods $\tilde{T}$ and T.

We therefore calculate the optimal willingness to pay $WTP^{OB}(\hat{\mu}_\tau, \tau)$ at relative time τ which the agent would commit to at time $t = 0$. To simplify our analysis and build on the results from the previous section, we assume that the decision-maker does not take the possibility of buying information into account when choosing her bias. This assumption seems appropriate when the probability of purchasing information, ϵ, is small.

Proposition 4. *Assume that an agent with positive belief utility chooses an optimal biased Bayesian updating process. Let the subjective belief at relative time τ be $0 < \hat{\mu}_\tau < 1$. The agent's willingness to pay evaluated at period 0, $WTP^{OB}(\hat{\mu}_\tau, \tau)$, satisfies*

$$\lim_{T\to\infty} WTP^{OB}(\hat{\mu}_\tau, \tau) = -\tilde{L}(\hat{\mu}_\tau) \tag{19}$$

where $\tilde{L}(\hat{\mu}) = (1-\tau)b(\hat{\mu}) - \int_0^{\hat{\mu}} c dG(c)$ is the per-period utility of a low type with belief utility $(1-\tau)b(\hat{\mu})$.

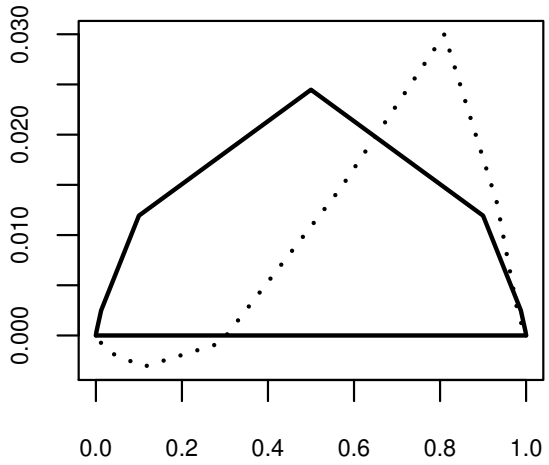
Proof. We know that high-type beliefs converge to 1 while low type beliefs stay close to μ_L^*. We also know that $\hat{\sigma}_T^L \to 0$ and $\hat{\sigma}_T^H \to 0$ and that there are constants $m_1, m_2 > 0$ such that $m_1 < \hat{\sigma}_T^L / \hat{\sigma}_T^H < m_2$. Hence, the probability at relative time τ that the agent is a low type provided that $\hat{\mu}_{\lfloor \tau T^j \rfloor} < 1$ converges to 1. Therefore, learning one's type decreases the agent's total utility to 0 with probability approaching 1 as $T \to \infty$ and destroys belief utility $(1-\tau)b(\hat{\mu}_\tau)$ (since low type logit-beliefs follow a driftless random walk with vanishing variance). □

Intuitively, an agent with subjective belief below 1 is asymptotically likely to be a low type, as otherwise her beliefs would have converged rapidly to 1. Proposition 2 implies that her beliefs in the low state follow a driftless random walk with vanishing variance and hence stay around $\hat{\mu}_\tau$. This implies that her belief utility over the remaining relative time $1-\tau$ is approximately $(1-\tau)b(\hat{\mu}_\tau)$. Buying information, on the other hand, would reveal her to be a low type immediately and yield a payoff of 0.

The economic significance of this result is that for low subjective beliefs $\hat{\mu}$ (and τ not too large) the optimal willingness to pay is negative, since the benefits of sustaining belief utility exceed the costs of mistaken choices, while for high subjective beliefs the optimal WTP is positive, since this relationship is reversed.[4] Thus Proposition 4 implies that, consistent with our empirical findings, the optimally biased agent will have a negative value of information when her belief is low and a positive value of information when her belief is high. This effect is mitigated for larger τ when belief utility is aggregated over fewer periods and hence becomes relatively less important; in this case information demands begin to resemble traditional, unbiased demands.

[4]Note, that $WTP^{OB}(\hat{\mu}_\tau, \tau)$ equals $-L(\hat{\mu}_\tau)$ for $\tau = 0$. Therefore, the biased Bayesian's willingness to pay for information is negative for low beliefs because $L(\mu_L^*) > 0$.

Figure S-2: Numerical optimum information demand functions for finite T and binary signals



Plots information values for realizable values of $\hat{\mu}_{[\tau T]}$ for $T = 31$, and $[\tau T] = 10$ for the unbiased Bayesian (solid lines) and agent with optimal simple updating bias (dotted lines) cases. The remaining parameters are fixed in both cases at $\mu_0 = 0.5$, $c \sim U[0, 1]$, $b(\hat{\mu}) = \frac{1}{4}\hat{\mu}$, $p = 0.75$, $q = 0.25$

Figure S-2 plots an example of the finite-T numerical demands generated by our model for both an unbiased and an optimally biased Bayesian. The unbiased Bayesian always values information positively, and values it most at intermediate beliefs where uncertainty is highest. The optimally biased agent, on the other hand, places a negative value on information for low levels of confidence and only assigns a positive value above a threshold level of confidence.

S-3 Additional Tables

Table S-4: Quiz Performance: Summary Statistics

		Correct		**Incorrect**		**Score**	
	N	Mean	SD	Mean	SD	Mean	SD
Overall							
Restricted Sample	656	10.2	4.3	2.7	2.1	7.4	4.8
Full Sample	1058	9.7	4.3	3.0	2.4	6.8	4.9
By Quiz Type							
1	79	8.1	3.1	1.7	1.2	6.4	3.3
2	85	13.0	2.9	2.7	2.1	10.3	3.4
3	69	8.9	3.3	3.0	2.1	5.9	3.8
4	74	12.2	3.8	3.1	2.3	9.2	4.6
5	75	6.5	1.6	4.0	2.3	2.5	2.8
6	63	14.5	4.5	2.3	1.7	12.3	4.7
7	73	7.6	2.6	2.2	1.7	5.4	3.1
8	69	13.6	2.8	3.2	1.8	10.4	3.3
9	69	7.3	3.5	2.7	2.8	4.7	4.5
By Gender							
Male	314	10.6	4.2	2.7	2.3	7.9	4.8
Female	342	9.7	4.4	2.8	2.0	6.9	4.8

Table S-5: Priors are Sufficient Statistics for Lagged Information: Full Sample

Regressor	Round 2	Round 3	Round 4
δ	1.070	0.938	0.906
	$(0.139)^{***}$	$(0.121)^{***}$	$(0.149)^{***}$
β_H	0.201	0.226	0.300
	$(0.026)^{***}$	$(0.030)^{***}$	$(0.041)^{***}$
β_L	0.133	0.205	0.251
	$(0.050)^{***}$	$(0.036)^{***}$	$(0.045)^{***}$
β_{-1}	-0.027	0.030	0.020
	(0.042)	(0.034)	(0.039)
β_{-2}		0.023	0.068
		(0.039)	(0.045)
β_{-3}			0.058
			(0.051)
N	999	999	999
R^2	-	-	-

Each column is a regression. The outcome in all regressions is the log posterior odds ratio. Reported coefficients are on the log prior odds ratio (δ), the log likelihood ratio for positive and negative signals (β_H and β_L), and the log likelihood ratio of the signal received τ periods earlier ($\beta_{-\tau}$). The estimation sample includes participants whose beliefs were always within $(0, 1)$ and who updated their beliefs at least once and never in the wrong direction. Estimation is via IV using the average score of other participants who took the same (randomly assigned) quiz as an instrument for the log prior odds ratio. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table S-6: Conservative and Asymmetric Belief Updating: Only non-zero updates

Regressor	Round 1	Round 2	Round 3	Round 4	All Rounds	Unrestricted
Panel A: OLS						
δ	0.777	0.946	0.943	1.009	0.937	0.888
	$(0.042)^{***}$	$(0.020)^{***}$	$(0.030)^{***}$	$(0.027)^{***}$	$(0.016)^{***}$	$(0.014)^{***}$
β_H	0.448	0.400	0.456	0.568	0.487	0.264
	$(0.021)^{***}$	$(0.020)^{***}$	$(0.024)^{***}$	$(0.035)^{***}$	$(0.016)^{***}$	$(0.013)^{***}$
β_L	0.477	0.422	0.457	0.471	0.454	0.211
	$(0.033)^{***}$	$(0.025)^{***}$	$(0.027)^{***}$	$(0.027)^{***}$	$(0.016)^{***}$	$(0.011)^{***}$
$\mathbb{P}(\beta_H = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_L = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_H = \beta_L)$	0.471	0.492	0.989	0.030	0.083	0.000
N	420	413	422	458	1713	3996
R^2	0.754	0.882	0.874	0.864	0.846	0.798
Panel B: IV						
δ	1.262	0.953	1.058	0.943	1.032	0.977
	$(0.325)^{***}$	$(0.098)^{***}$	$(0.141)^{***}$	$(0.157)^{***}$	$(0.078)^{***}$	$(0.060)^{***}$
β_H	0.617	0.401	0.456	0.578	0.496	0.273
	$(0.129)^{***}$	$(0.024)^{***}$	$(0.025)^{***}$	$(0.041)^{***}$	$(0.017)^{***}$	$(0.013)^{***}$
β_L	0.414	0.421	0.450	0.477	0.446	0.174
	$(0.052)^{***}$	$(0.024)^{***}$	$(0.029)^{***}$	$(0.034)^{***}$	$(0.015)^{***}$	$(0.027)^{***}$
$\mathbb{P}(\beta_H = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_L = 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\mathbb{P}(\beta_H = \beta_L)$	0.233	0.556	0.866	0.035	0.049	0.004
First Stage F-statistic	4.85	14.47	11.24	8.40	14.86	20.61
N	420	413	422	458	1713	3996
R^2	-	-	-	-	-	-

Notes:

1. Each column in each panel is a regression. The outcome in all regressions is the log posterior odds ratio. δ is the coefficient on the log prior odds ratio; β_H and β_L are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. Bayesian updating (for both biased and unbiased Bayesians) corresponds to $\delta = \beta_H = \beta_L = 1$.
2. Estimation samples are restricted to participants whose beliefs were always within $(0, 1)$. Columns 1-5 further restrict to participants who updated their beliefs in every round and never in the wrong direction; Column 6 includes participants violating this condition. Columns 1-4 examine updating in each round separately, while Columns 5-6 pool the 4 rounds of updating.
3. Estimation is via OLS in Panel A and via IV in Panel B, using the average score of other participants who took the same (randomly assigned) quiz variety as an instrument for the log prior odds ratio.
4. Heteroskedasticity-robust standard errors in parenthesis; those in the last two columns are clustered by individual. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table S-7: Updating is not Differential by Prior

Regressor	Round 1	Round 2	Round 3	Round 4	All Rounds	Unrestricted
Panel A: OLS						
δ	0.908	0.944	0.952	0.958	0.944	0.885
	(0.029)***	(0.019)***	(0.027)***	(0.022)***	(0.012)***	(0.014)***
δ_H	-0.150	-0.040	-0.020	0.058	-0.037	0.006
	(0.053)***	(0.030)	(0.046)	(0.046)	(0.023)	(0.026)
β_H	0.361	0.295	0.334	0.434	0.369	0.264
	(0.018)***	(0.017)***	(0.021)***	(0.030)***	(0.013)***	(0.013)***
β_L	0.268	0.270	0.302	0.354	0.298	0.212
	(0.026)***	(0.020)***	(0.022)***	(0.024)***	(0.012)***	(0.011)***
N	612	612	612	612	2448	3996
R^2	0.808	0.891	0.875	0.860	0.854	0.798
Panel B: IV						
δ	0.876	1.070	1.398	0.830	1.071	0.976
	(0.518)*	(0.189)***	(0.257)***	(0.158)***	(0.108)***	(0.096)***
δ_H	0.092	-0.287	-0.544	0.166	-0.167	0.002
	(0.539)	(0.220)	(0.307)*	(0.237)	(0.132)	(0.123)
β_H	0.409	0.292	0.335	0.437	0.369	0.273
	(0.045)***	(0.017)***	(0.021)***	(0.037)***	(0.013)***	(0.014)***
β_L	0.277	0.243	0.216	0.385	0.268	0.175
	(0.151)*	(0.044)***	(0.062)***	(0.049)***	(0.027)***	(0.041)***
N	612	612	612	612	2448	3996
R^2	-	-	-	-	-	-

Notes:

1. Each column in each panel is a regression. The outcome in all regressions is the log posterior odds ratio. δ is the coefficient on the log prior odds ratio; δ_H is the coefficient on an interaction between the log prior odds ratio and an indicator for a positive signal; β_H and β_L are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. Bayesian updating corresponds to $\delta = \beta_H = \beta_L = 1$ and $\delta_H = 0$.
2. Estimation samples are restricted to participants whose beliefs were always within $(0, 1)$. Columns 1-5 further restrict to participants who updated their beliefs at least once and never in the wrong direction; Column 6 includes participants violating this condition. Columns 1-4 examine updating in each round separately, while Columns 5-6 pool the 4 rounds of updating.
3. Estimation is via OLS in Panel A and via IV in Panel B, using the average score of other participants who took the same (randomly assigned) quiz variety as an instrument for the log prior odds ratio.
4. Heteroskedasticity-robust standard errors in parenthesis; those in the last two columns are clustered by individual. Statistical significance is denoted as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.