Managing Self-Confidence:
Theory and Experimental Evidence*

Markus M. Möbius
Iowa State University and NBER

Muriel Niederle
Stanford University and NBER

Paul Niehaus
UC San Diego

Tanya S. Rosenblat
Iowa State University

October 21, 2010

Abstract

A large literature in social psychology suggests that agents process information about
their own ability in a biased manner. While this finding has generated exciting research in
behavioral economics, some of the fundamental evidence has also been criticized as being
potentially consistent with Bayesian updating. We reevaluate the evidence from social psy-
chology by conducting a large-scale experiment with 656 undergraduate students. Subjects
perform an IQ-test and receive noisy feedback on their performance while we track the evo-
lution of their beliefs. Our experimental design allows us to precisely measure the relevant
belief distribution of subjects simply and repeatedly. Our experiment confirms that sub-
jects are asymmetric and put greater weight on positive performance signals compared to
negative signals. However, we also find that subjects are conservative and update too little
in response to both positive and negative signals. Subjects appear much more Bayesian
when receiving feedback about someone else’s performance. We show that these seemingly
related phenomena naturally arise in a simple model of biased information processing in-
spired by the anticipatory utility model of Brunnermeier and Parker (2005). Our model
is parsimonious and adds just a single parameter to the standard model. We also find
that women are more conservative than men, but not more asymmetric. High-confidence
women have a higher value for information than men.

JEL Classification: C91, C93, D83

Keywords: asymmetric belief updating, conservatism, information aversion

*We are grateful to Nageeb Ali, Roland Benabou, Gary Chamberlain, Rachel Croson, Gordon Dahl, John
List, Al Roth and Lise Vesterlund for helpful discussions. We would like to thank seminar participants at
University of Chicago, Iowa State University, Federal Reserve Bank of Boston, the Institute for Advanced
Study (Princeton), Experimental Economics Conference (UCSB), Workshop in Behavioral Public Economics
(Insbruck), and 2009 North American Meetings of the Economic Science Association for their feedback. 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

Standard economic theory assumes that rational agents process information about their own ability like dispassionate Bayesians. Social psychologists have questioned this assumption by pointing out that people systematically rate their own ability as “above average”. To take one classic and widely-cited example, 88% of U.S. drivers consider themselves safer than the median driver (Svenson 1981). A quickly expanding literature in behavioral economics (Koszegi 2006, Van den Steen 2004) and finance (Malmendier and Tate 2008, Barber and Odean 2001) has explored the implications of overconfidence on economic decision-making.

At the same time, economists have pointed out that much of the commonly cited evidence on biased information processing is in fact consistent with fully rational information processing. Zábojník (2004) and Benoit and Dubra (2008) have shown that Bayesian updating can quite easily generate highly skewed belief distributions. For example, if there are equally many safe and unsafe drivers and only unsafe drivers have accidents, then a majority of drivers will always rate themselves as safe with a higher than 50 percent probability (namely the good drivers and the bad drivers who did not have an accident yet). People might also disagree on the definition of what constitutes a safe driver (Santos-Pinto and Sobel 2005) or even have heterogeneous priors (Van den Steen 2004).

Our paper makes two contributions. First, we provide a theoretical foundation for tests of self-confidence management. We build on Brunnermeier and Parker’s (2005) (henceforth BP) concept of anticipatory utility and show that it gives rise not only to ego-biased belief updating but also to additional measurable behaviors not previously associated with biased updating. This is important since it lets us test a full array of theoretical predictions, rather than search for an isolated bias. Second, we implement these tests in a carefully controlled experimental environment using a design that is robust to the critiques discussed above. We elicit subjects’ beliefs about well-defined events in an incentive-compatible manner and study their evolution, opening up the “black box” of updating.

The model describes an agent who has either high or low ability. He will eventually have to choose whether or not to take an action whose payoff is positive only if his type is high, so he places an instrumental value on information. However, he also derives anticipatory utility from his beliefs about his future payoff, which implies that he prefers to believe that he is the high type. Following BP we suppose that the agent decides at an initial stage how to interpret the informativeness of signals and how to value information, taking into account the competing demands of decision-making and anticipatory utility. The model is parsimonious since it adds just a single parameter to the standard model.

---

1See Englmaier (2006) or Benoit and Dubra (2008) for overviews of the evidence on over-confidence.
Like other behavioral models ours can explain why agents are asymmetric, putting greater weight on positive information about their own ability than on negative information. What sets our model apart, however, is that it reveals close connections between asymmetry and other apparently anomalous behaviors. It predicts that agents should be conservative, responding less than a Bayesian to information. The intuition is that while asymmetry alone can help an agent maintain a high belief in her ability even if she is low-skilled it also increases the probability that the agent takes the wrong action eventually. By also becoming more conservative, the biased Bayesian can reduce the variance of her belief distribution in the bad state of the world. Conservatism complements asymmetry in helping the agent maintain a high belief in her ability in the bad state of the world while minimizing the chance of taking the wrong action in the end. The model also predicts that less confidence agents may be information-averse, willing to pay to avoid learning their types, while more confident agents have a greater demand for information than a perfect Bayesian.

We conduct a large-scale experiment with 656 undergraduate students. Our design addresses the critiques of economists by carefully measuring subjects’ prior and posterior beliefs. In our experiment, subjects first perform an IQ-test and we then measure their initial belief that they are among the top half of performers. We then provide each subject with four sequential and independent binary signals on their performance. Each signal tells tells them whether they are among the top or bottom half of performers and each signal is correct with 75 percent probability. After each signal, we again elicit a subject’s belief to be among the top half of performers. A novel and convenient feature of our design is that it allows us to summarize the belief distribution in a single number, namely the probability of being among the top half of performers.

Our experimental design also introduces two important methodological innovation. (1) We elicit beliefs by asking subjects what type of binary lottery would make them indifferent between the diagnostic lottery and the lottery we are trying to measure (being among the top half). Unlike the often used quadratic scoring mechanism, this new incentive-compatible mechanism is robust to subjects being risk averse. The mechanism can even accommodate more non-standard utility models: the only assume that subjects’ preferences are monotonic such that lotteries that give the high payoff with higher probability are preferred. Our mechanism has also been independently discovered by Karni (2009). (2) By focusing on the probability of scoring among the top half of subjects, we drastically simplify belief elicitation because the entire belief distribution is summarized in a single number. This allows us to measure beliefs repeatedly without exhausting subjects.

We estimate empirical specifications of belief updating that nest Bayes rule and our own model. Consistent with both, we find that information is persistent in the sense that subjects
priors are fully incorporated into their posteriors. Consistent only with the latter, we find that subjects are both conservative and asymmetric updaters. On average our subjects revise their beliefs by only 35% as much as Bayesians with the same priors would. Moreover, subjects who receive positive feedback revise their beliefs by 15% more on average than those who receive negative feedback. The most transparent manifestation of asymmetry is that subjects who received two positive and two negative signals – and thus learned nothing – end up significantly more confident than they began. We take this as unambiguous evidence of self-serving bias.

An important question about these results is whether they reflect motivated behavior or cognitive limitations. It is widely recognized that Bayesian updating is an imperfect positive model even when self-confidence is not at stake. To distinguish these interpretations we conduct a placebo experiment, structurally identical to our initial experiment except that subjects report their beliefs about the performance of a “robot” rather than their own performance. Belief updating in this second experiment is significantly and substantially closer to Bayesian, implying that the desire to manage self-confidence is an important driver of updating behavior.

To measure the demand for information we conclude our experiment by allowing subjects to bid for noiseless information on their relative performance. We then test the null hypothesis that subjects’ valuations for feedback are weakly positive. This must hold if subjects use information purely to improve their decision-making. However, our model predicts agents may be averse to feedback. We find that approximately 10% of our subjects are information-averse, willing to pay to avoid learning their type. We also show that this result cannot be explained by plausible forms of measurement error in reported willingness to pay.

Our results provide support for recent theoretical work that has modeled the potential benefits of a high self-confidence. Self-confidence may directly enhance well-being (Akerlof and Dickens 1982, Caplin and Leahy 2001, Brunnermeier and Parker 2005, Koszegi 2006), may compensate for limited self-control, (Brocas and Carrillo 2000, Benabou and Tirole 2002) or may directly enhance performance (Comppte and Postlewaite 2004). These theories differ in their assumptions about how people are able to manage their self-confidence, with some

---

2 As we discuss below, psychological research on belief updating suffers from many of the same problems that haunt work on over-confidence. Though many experiments purport to demonstrate an “attribution bias” whereby agents selectively interpret information in a self-serving manner (Langer and Roth 1975, Miller and Ross 1975, Kelley and Michela 1980), much of this behavior is also consistent with Bayesian updating (Ajzen and Fishbein 1975, Wetzel 1982). Further ambiguities arise because beliefs are not formally elicited and because in many cases subjects are not given the details of the data-generating process.

3 A large literature in psychology during the 1960s tested Bayes rule for ego-independent problems such as predicting which urn a series of balls were drawn from; 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 subjects update. Rabin and Schrag (1999) and Rabin (2002) study the theoretical implications of specific cognitive forecasting and updating biases.
emphasizing non-Bayesian updating while others emphasize strategic information acquisition. Our results suggest that both mechanisms are relevant.

Our findings also contribute to research on gender differences in confidence. A large literature in psychology and a growing one in economics have emphasized that men tend to be more (over-) confident than women and that this has important implications for their economic outcomes. In principal this could arise because of differences in the way men and women acquire information or differences in the way they process it. We find empirical support only for the former channel: women are not significantly less asymmetric than men, but high-confidence women have differentially higher demand for feedback. This implies that over time over-confidence will become concentrated among men. Encouragingly, it also implies that a planner could potentially influence the gender gap in confidence simply by providing targeted feedback to under-confident women.

The rest of the paper is organized as follows. Section 2 develops the model. Section 3 describes the details of our experimental design and Section 4 summarizes the experimental data. Section 5 discusses econometric methods and presents results for belief updating dynamics, and Section 6 presents results on information acquisition behavior. Section 7 discusses gender differences in both information processing and acquisition. Section 8 concludes.

2 Theory

We consider an agent who can either be of high type \( H \) or low type \( L \). There are \( T \) discrete time periods and the agent can observe i.i.d. binary signals about her type in each period. In the last period the agent has to decide whether to make an investment at cost \( c \) which pays 1 if she is of high type and 0 otherwise. Not investing gives utility 0. The investment provides the agent with incentives to form accurate beliefs about her own ability. The investment cost \( c \) is drawn in the final period from a uniform distribution over the interval \([c, 1]\) with \( c > 0 \). The timeline of the model is shown in figure 1.

![Timeline of model](image-url)
We first analyze this model under the assumption that the agent is a “perfect Bayesian” who uses the correct signal distribution when applying Bayes rule to form a posterior. We then examine the information processing of an “optimally biased Bayesian” who also uses Bayes rule but can choose at time $t = 0$ how to interpret the informativeness of positive and negative signals. The biased Bayesian derives anticipatory utility during the first $T$ periods from believing that she is a high-type agent. Biased information processing can increase anticipatory utility at the cost of being more likely to take the wrong action in the last period.

2.1 Information Processing of a “Perfect Bayesian”

The agent has a subjective prior belief $\mu \in (0, 1)$ that she is a high type and in each period $t = 1, \ldots, T$ receives a binary signal $s_t \in \{H, L\}$ about her type. The signals are conditionally independently distributed: a high type agent receives a high signal with probability $p > \frac{1}{2}$ and a low type agent receives a high signal with probability $q < \frac{1}{2}$. The perfect Bayesian derives her posterior $\mu^t$ using Bayes rule. In the final period the agent will invest if she has a sufficiently high belief to be a high type that is $\mu^T > c$. We assume throughout that $\mu < c$. This implies that the agent has demand for learning before deciding about the costly investment.

We denote the number of $H$ signals received by time $t$ with $S^t_H$ and the number of $L$ signals with $S^t_L$. We can calculate the Bayesian posterior $\mu^t$ using the logistic function $\logit(x) = \log(x/(1 - x))$ and the log-likelihood ratios for high and low signals:

$$\logit(\mu^t) = \logit(\mu) + S^t_H \log \left(\frac{p}{q}\right) + S^t_L \log \left(\frac{1 - p}{1 - q}\right)$$

We will refer to $\lambda_H = \log \left(\frac{p}{q}\right)$ and $\lambda_L = \log \left(\frac{1 - p}{1 - q}\right)$ as the informativeness of a positive and negative signal, respectively. The vector $\vec{\lambda} = (\lambda_H, \lambda_L)$ summarizes the signal structure of the game.

Logit-beliefs evolve as a random walk with a drift that depends on the agent’s type. The mean logit-belief $\gamma^t_L = E \left( \logit(\mu^t) | L \right)$ of a low type agent is linearly decreasing over time:

$$\gamma^t_L = \logit(\mu) + t [q\lambda_H + (1 - q)\lambda_L]$$

In contrast, the mean logit-belief of a high-type agent evolves with a positive drift and its mean $\gamma^t_H = E \left( \logit(\mu^t) | H \right)$ is linearly increasing over time:

$$\gamma^t_H = \logit(\mu) + t [p\lambda_H + (1 - p)\lambda_L]$$

\footnote{We adopt the tie-breaking convention that the agent does not invest if she is indifferent between investing and not investing.}
Figure 2 graphs the distribution of beliefs for high and low type agents. The two highlighted lines show the evolving mean logit-beliefs of low and high types. The two curves indicate that the distribution of logit-beliefs in the final period $T$ will be approximately normally distributed for larger $T$ with standard deviation $\sigma_T^L$ and $\sigma_T^H$:

$$\sigma_T^{L^2} = T \left[ q \lambda_H^2 + (1-q)\gamma_L^2 - \gamma_L^2 \right]$$

$$\sigma_T^{H^2} = T \left[ p \lambda_H^2 + (1-p)\gamma_H^2 - \gamma_H^2 \right]$$

It will be useful for our analysis of the “biased Bayesian” agent to graphically understand the investment decision of the perfect Bayesian for large $T$. The agent will invest at time $T$ if and only if her logit-belief is greater than the realized cutoff logit($c$). As the agent accumulates more and more signals the mean logit-beliefs of the low and high type agent converge to minus and plus infinity at rate $T$, respectively. At the same time, the standard deviation increases only at rate $\sqrt{T}$ in both cases. Therefore, the agent will make fewer and fewer investment mistakes as $T \to \infty$ because the probability that her logit-belief is below or above the cutoff in both states of the world converges to 1. The expected utility of the low and high type agent will converge to 0 and $1 - c$.

**Proposition 1** The expected utility of a perfect Bayesian decision-maker is $\mu(1 - E(c)) + O(\exp(-aT))$ for some constant $a > 0$. 

7
2.2 Information Processing of a “Biased Bayesian”

The “biased Bayesian” differs in two dimensions from a perfect Bayesian decision maker. First, the biased Bayesian derives direct utility from beliefs. In particular, we build on the model of Brunnermeier and Parker (2005) and assume that the agent derives anticipatory utility $\hat{\mu}^t(1 - E(c))$ in periods $1 \leq t \leq T$. Note, that we define anticipatory utility simple as expected utility of a perfect Bayesian who has access to an infinite number of signals and therefore eventually fully learns her type (see proposition 1).

Second, we allow a biased agent two parameters with which to deterministically bias their beliefs as they receive new information. This allows us to take the model to the data, and estimate those parameters. A biased Bayesian chooses at time $t = 0$ the informativeness of positive and negative signals, $\hat{\lambda}_H > 0$ and $\hat{\lambda}_L < 0$. The vector $\hat{\lambda} = (\hat{\lambda}_H, \hat{\lambda}_L)$ summarizes how the biased Bayesian interprets the signal structure.

We define the parameters $\beta_H = \frac{\hat{\lambda}_H}{\lambda_H}$ and $\beta_L = \frac{\hat{\lambda}_L}{\lambda_L}$ as the decision-maker’s relative responsiveness to negative and positive information. The biased Bayesian derives his posterior belief $\hat{\mu}^t$ by using Bayes rule just as before:

$$\logit(\hat{\mu}^t) = \logit(\mu) + S^t_H \hat{\lambda}_H + S^t_L \hat{\lambda}_L$$  \hspace{1cm} (4)

We denote the mean logit beliefs of the low and high type biased Bayesian with $\gamma^t_L$ and $\gamma^t_H$ and the standard deviations with $\sigma^t_L$ and $\sigma^t_H$.

Following Brunnermeier and Parker (2005), we write the total utility of the decision-maker as the sum of anticipatory utility and realized utility from actual investment realized at time $T + 1$:

$$U(\hat{\lambda}|\alpha, \mu, \bar{\lambda}) = E_{\hat{\mu}^t} \left[ \alpha \frac{1}{T} \sum_{t=1}^{T} \hat{\mu}^t(1 - E(c)) + E_c \left( \max(\hat{\mu}^T - c, 0) \right) \right]$$  \hspace{1cm} (5)

The parameter $\alpha$ captures the relative importance of anticipatory utility.\(^5\)

---

\(^5\)In Brunnermeier and Parker’s (2005) original model anticipatory utility at time $t$ is the decision-maker’s expected utility assuming that she receives $T - t$ signals before making an investment. We modified their specification for the sake of analytical tractability.

\(^6\)Formally, these expressions are defined through equations 2, 3 and 4 where we replace $\lambda_H$ and $\lambda_L$ with $\hat{\lambda}_H$ and $\hat{\lambda}_L$.

\(^7\)Note, that the outer expectation is taken over all sample paths $\hat{\mu}^t$. Each sample path is generated according to formula 4 for a particular realization of signals $(S_H, S_L)$. Importantly, these signals are generated by the correct data generating process described by $\mu$ and $\bar{\lambda}$. 

---

All proofs are delegated to the Appendix.
The optimally biased Bayesian chooses her responsiveness to positive and negative information in order to maximize her utility. Note, that anticipatory utility is necessary to induce biased information processing.

**Proposition 2** A biased Bayesian who derives no utility from anticipation ($\alpha = 0$) processes information like a perfect Bayesian and chooses $\beta_H = \beta_L = 1$ for any $T > \mathcal{T}(\xi, p, q)$.

The analysis of the biased Bayesian becomes interesting when anticipatory utility matters ($\alpha > 0$) but does not overwhelm the agent’s realized utility from investment. When the parameter $\alpha$ becomes too large, the agent will simply bias beliefs in such a way that she always converges to high beliefs regardless of her type. The following assumption provides us with a suitable upper limit on $\alpha$ for our purposes:

**Assumption 1** (*Long-Run Learning*)

$$\alpha < \frac{1}{2} \frac{E(c)}{1 - E(c)} \quad (6)$$

The assumption ensures that the agent prefers to stay unbiased rather than adopt biased beliefs which make her take the wrong action with probability $\frac{1}{2}$ for any $c$ in the low state of the world. Formally, biased information processing creates additional expected anticipatory utility of at most $\alpha(1 - \mu)(1 - E(c))$. At the same time, if the low-type agent makes an incorrect investment decision with probability of $\frac{1}{2}$ for any $c$, then she incurs an additional expected loss of $\frac{1}{2}(1 - \mu)E(c)$.

### 2.3 Optimally Biased Bayesians

The following definition will be useful to classify optimal Bayesian bias in our model.

**Definition 1** We say that an biased Bayesian is conservative if the agent’s relative responsiveness to both positive and negative information is less than 1. We say that the agent has an asymmetric bias if her relative responsiveness to positive information is greater than her relative responsiveness to negative information.

Much of the intuition for optimal Bayesian bias can be derived from the following special case ($\frac{1}{2} < \theta < 1$):

$$\hat{\lambda}_H = T^{-\theta} \lambda_H$$

$$\hat{\lambda}_L = -T^{-\theta} \frac{q}{1 - q} \lambda_H \quad (7)$$
Note, that this agent is conservative since her responsiveness to positive and negative information tends to 0 as $T \to \infty$. The agent is also asymmetric: in fact her logit belief follows a driftless random walk if she is a low type ($\tilde{\gamma}_T = 0$). Because of this property we will refer to this particular bias from now on as the *downward neutral bias*.

Her final logit-belief of the low type is therefore approximately normally distributed with mean logit($\mu$) and standard deviation $O\left(T^{1/2-\theta}\right)$ which tends to zero as $T \to \infty$. Therefore, we can make the probability that the low-type invests in the last period arbitrarily small by choosing $T$ large enough. At the same time the high-type agent’s logit-belief has a positive drift and her mean logit belief will be of order $O\left(T^{1-\theta}\right)$ which tends to infinity as $T \to \infty$. The standard deviation of her logit-belief, on the other hand, will be again of order $O\left(T^{1/2-\theta}\right)$. Hence, the high-type agent will take the correct action in the final period with probability approaching 1. The anticipatory utility of the agent will be approximately $\alpha(1 - E(c))(\mu^2 + 1 - \mu)$ because the high type agent will believe that she is of type most of the time, while the low-type agent will hold on to her belief that is high-type with probability $\mu$. Note, that the anticipatory utility of the biased Bayesian is strictly larger than the anticipatory utility of the perfect Bayesian.

The special case of downward neutral bias provides a lower bound on the utility of the optimally biased Bayesian. It shows that the biased Bayesian can always strictly increase her anticipatory utility while keeping the probability of mistakes in the final period arbitrarily small for sufficiently large $T$. The special case of downward neutral bias exhibits both conservatism and asymmetry. We now show that these two properties characterize the bias of the optimal Bayesian as well.

**Theorem 1** The responsiveness of the optimally biased Bayesian to both positive and negative information tends to zero as $T \to \infty$. Therefore, all optimally biased Bayesians are conservative for sufficiently large $T$. Moreover, the optimally biased Bayesian responds more to both positive than negative information as $T \to \infty$. Therefore, all optimally biased Bayesians are asymmetric for sufficiently large $T$.

The intuition for conservatism follows readily from our long-run learning assumption: a biased low-type agent who maintains her self-confidence and is *not* conservative will have highly uninformative beliefs and will essentially invest randomly with probability $\frac{1}{2}$ in the final period. The potential gain in anticipatory utility, however, cannot outweigh the loss from taking the wrong action in the final period under assumption [1].

We can show that the optimally biased Bayesian behaves “almost” like an agent with downward neutral bias for large $T$. 

10
Plots responsiveness to positive and negative signals ($\beta_H$ and $\beta_L$) for various values of $T$ and for $\mu = \frac{1}{2}$, $c = 0.9$, $p = \frac{3}{4}$, $q = \frac{1}{4}$, and $\alpha = 0.2$.

**Proposition 3** For the optimally biased Bayesian, the ratio of the agent’s responsiveness to positive versus negative information converges to $\frac{1 - q}{q}$ as $T \to \infty$.

Figure 3 shows the results of a numerical simulation for the case $\alpha = 0.2$, $\mu = \frac{1}{2}$ and $c = 0.9$ over the range $10 < T < 40$. Signals are symmetric and each signal is correct with probability 0.75. The simulations show that the decision-maker rapidly becomes both conservative and asymmetric for finite $T$. As predicted, the ratio $\beta_H/\beta_L$ converges to 3 - this implies that for large $T$ the decision-maker behaves like an agent with downward neutral bias.

Intuition suggests that agents who place relatively more importance on the anticipatory component of their utility will be relatively more conservative and asymmetric. This claim certainly holds when we compare a fully rational agent ($\alpha = 0$) to an agent with anticipatory utility $\alpha > 0$. Intermediate cases can be examined numerically. In figure 4 we set $T = 21$ and vary $\alpha$ from 0 to 2 while keeping other parameters the same as in figure 3. The average responsiveness to information, $(\beta_H + \beta_L)/2$, decreases with $\alpha$, while the ratio $\beta_H/\beta_L$ increases $^8$

$^8$The quantity $\beta_H - \beta_L$ does not vary monotonically with $\alpha$, as it measures both responsiveness and asymmetry.
Figure 4: Conservatism and Asymmetry Increase with $\alpha$

Plots conservatism $(1 - (\beta_H + \beta_L)/2)$ and asymmetry $(\beta_H/\beta_L)$ for different values of the preference parameter $\alpha$. Other parameters are fixed at $\mu = \frac{1}{2}, c = 0.9, p = \frac{3}{4}, q = \frac{1}{4}$, and $T = 21$.

2.4 Value of Information

We now analyze how biased agents value information. Assume, that the agent is presented at time $[\tau T]$ with an opportunity to purchase a perfectly informative signal. Here, $0 \leq \tau < 1$ denotes the relative time when the agent can buy information and $[\tau T]$ is the closest rounded up or rounded down integer time. For example, $\tau = \frac{1}{2}$ indicates that the agent can buy information midway through the learning process. We are interested in the agent’s willingness to pay, $WTP(x, \tau | \alpha, \mu, \bar{\lambda}, \bar{\lambda})$, where $x$ denotes a feasible belief of the agent at relative time $\tau$.

Consistent with our modeling approach we assume that the decision-maker chooses her willingness to pay at time 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 is appropriate when the probability of purchasing information is small.

To formally derive the WTP, we start with the biased Bayesian’s utility at relative time $\tau$.

---

9Note, that the biased Bayesian can only reach at most $[\tau T] + 1$ distinct beliefs at relative time $\tau$ when her initial belief is $\mu$. 

---
if she declines to purchase information:

\[ U^{\text{noinfo}}(x, \tau | \alpha, \vec{\lambda}, \vec{\tilde{\lambda}}) = E_{\hat{\mu}^T} \left[ \begin{array}{c}
\frac{1}{T} \sum_{t=\lfloor \tau T \rfloor +1}^{T} \hat{\mu}^t (1 - E(c)) + E_c (\max (\hat{\mu}^T - c, 0)) \\
\hat{\mu}^{\lfloor \tau T \rfloor} = x
\end{array} \right] \] (8)

We take the expectation over all sample paths \( \hat{\mu}^t \) such that \( \hat{\mu}^{\lfloor \tau T \rfloor} = x \). Note, that anticipatory utility becomes relatively less important than the utility from investing as \( \tau \) increases because we sum over only \( T - \lfloor \tau T \rfloor \) periods. Next, we derive the agent’s utility at time \( \tau \) if she purchases a perfect signal:

\[ U^{\text{info}}(x, \tau | \alpha, \vec{\lambda}, \vec{\tilde{\lambda}}) = E_{\hat{\mu}^T} \left[ \mu^{\lfloor \tau T \rfloor} (\alpha + 1) (1 - E(c)) \right] \] (9)

We can now formally define the agent’s willingness to pay for information:

\[ WTP(x, \tau | \alpha, \mu, \vec{\tilde{\lambda}}) = U^{\text{info}}(x, \tau | \alpha, \vec{\lambda}, \vec{\tilde{\lambda}}) - U^{\text{noinfo}}(x, \tau | \alpha, \vec{\lambda}, \vec{\tilde{\lambda}}) \] (10)

In the special case, where the agent is a perfect Bayesian (\( \alpha = 0 \)) and takes an action immediately (\( \tau = T \)) the above expression reduces to the short-term willingness to pay \( WTP^S(x) \):

\[ WTP^S(x) = x (1 - E(c)) - \int_{\xi}^{x} \frac{\max(0, x - c)}{1 - \xi} dc \] (11)

It is easy to see that \( WTP^S(x) \) is uniquely maximized at \( x = \xi + \frac{(1 - \xi)^2}{2} \) and is zero when \( x = 0 \) or \( x = 1 \): a Bayesian decision-maker values information the least when she is sure about her ability. Moreover, the short-term willingness to pay for information is never negative.

Our first result looks at the willingness to pay for information of the perfect Bayesian:

**Lemma 1** Consider some belief \( 0 < x < 1 \) and relative time \( 0 < \tau < 1 - \epsilon \) for some \( \epsilon > 0 \). The perfect Bayesian’s willingness to pay, \( WTP(x, \tau | 0, \vec{\lambda}, \vec{\tilde{\lambda}}) \), converges to zero as \( T \to \infty \).

The intuition for this result is simply that with enough periods to go and imprecise beliefs \( 0 < x < 1 \), the perfect Bayesian will accumulate sufficient information to take the correct action with probability approaching 1. Hence, the value of a perfect signal converges to zero.

In order to analyze the biased Bayesian’s willingness to pay, we introduce the function \( \gamma^{*}_L(x, \tau | \mu) \):

\[ \gamma^{*}_L(x, \tau | \mu) = \logit(x) + \tau (\logit(\xi) - \logit(\mu)) \] (12)

Note, that \( \gamma^{*}_L(\mu, \tau | \mu) \) describes the mean logit belief of a low-type agent whose bias is such that in the final period her mean logit belief of being a high type is just \( \logit(\xi) \). We also

\[ \text{Note, that each sample path } \hat{\mu}^t \text{ uniquely determines } \mu_t. \]
introduce the function $A^*_L((x, \tau|\mu))$:

$$A^*_L(x, \tau|\mu) = (1 - E(c)) \int_{\tau}^{1} \text{logit}^{-1}(\gamma^*_L(x, \tau'|\mu)) d\tau'$$ (13)

where $\text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ denotes the inverse logit function. This is the remaining expected anticipatory utility enjoyed by a low type at time $\tau$ with belief $x$ whose logit belief evolves exactly according to function $\gamma^*_L(x, \tau|\mu)$. Note, that due to our assumption $\mu < c$ we have $A^*_L(x, \tau|\mu) \geq (1 - \tau)x(1 - E(c))$.

**Theorem 2** Consider some biased Bayesian who places weight $\alpha > 0$ on anticipatory utility. Assume $0 < x < 1 - \delta$, $\mu \geq \frac{1}{2}$ and relative time $\epsilon < \tau < 1 - \epsilon$ for some $\epsilon, \delta > 0$. The agent’s willingness to pay satisfies:

$$\lim_{T \to \infty} WTP(x, \tau|\alpha, \mu, \vec{\lambda}, \vec{\hat{\lambda}}) = -\alpha A^*_L(x, \tau|\mu) + \frac{\text{logit}^{-1}(\gamma^*_L(x, T - \tau|\mu))^2 - c^2}{2(1 - \zeta)}$$ (14)

Intuitively, when beliefs are low, such that $x < \text{logit}^{-1}(\gamma^*_L(\mu, \tau|\mu))$, the agent has a negative willingness to pay: period 0 self knows that period $\tau$ self is a low-type because otherwise her beliefs would have rapidly converged to 1. Moreover, the agent’s belief is low enough to ensure that her final belief is likely going to be below $c$. Hence, the agent is unlikely to make a wrong investment. Learning one’s ability therefore only destroys belief utility and has no upside. When beliefs are higher, such that $\text{logit}^{-1}(\gamma^*_L(\mu, \tau|\mu)) < x < 1 - \delta$, period 0 self still knows that she is likely a low type and that she only ended up with a relatively high belief because of a lucky string of signal draws in the first $[\tau T]$ periods. Her likely final belief will be $\text{logit}^{-1}(\gamma^*_L(x, T - \tau|\mu)) > c$ which implies that she will wrongly invest with probability $(\text{logit}^{-1}(\gamma^*_L(x, T - \tau|\mu)) - c)/(1 - c)$. For sufficiently large $x$ the incentive to avoid mistakes will dominate the disutility from destroying anticipatory utility due to our long-run learning assumption \[\square\] and the willingness to pay for information becomes positive. To summarize, the biased Bayesian will tend to have a negative value of information when her belief is low and a positive value of information when her belief is high.

We conclude this section by comparing period 0’s willingness to pay for information at relative time $\tau$ with the naive biased Bayesian’s willingness to pay, who convinced her that the signal structure is $\vec{\lambda}$ instead of $\vec{\hat{\lambda}}$ but who is unaware of her bias and thinks of herself as a perfect Bayesian ($\alpha = 0$). We therefore define:

$$F(x, \tau|\alpha, \mu, \vec{\lambda}, \vec{\hat{\lambda}}) = \frac{WTP(x, \tau|0, \mu, \vec{\lambda}, \vec{\hat{\lambda}})}{\text{naive biased Bayesian’s WTP}} - \frac{WTP(x, \tau|\alpha, \mu, \vec{\lambda}, \vec{\hat{\lambda}})}{\text{Period 0’s WTP}}$$ (15)
The function $F()$ describe the agent’s feedback aversion. The absolute value of this effect measures the mental effort cost that period 0-self imposes on future selves’ demand for information. Positive values indicate that period 0 dislikes information more than the naive Bayesian while negative values indicate a taste for information. Our notion of feedback aversion captures the idea that people often avoid feedback on an unconscious level or discount the accuracy of costly feedback to justify avoidance (see Caplin and Leahy (2001) for examples).

**Proposition 4** Under the assumptions of theorem 2 the naive Bayesian’s willingness to pay satisfies:

$$\lim_{T \to \infty} WTP(x, \tau|0, \mu, \lambda, \lambda) = WTP^S(x)$$

Intuitively, the optimally biased agent updates her beliefs so slowly, that she does not expect to learn much until taking an action. Hence, the Bayesian component of her willingness to pay converges to a Bayesian’s immediate value of information. Consequently, feedback aversion will be strong both for low beliefs (when period 0 self imposes a dislike for information on her future self) and high beliefs (when period 0 self imposes a taste for information on her future self).

### 2.5 Discussion of Modeling Assumptions

**Anticipatory utility.** A growing literature in behavioral economics has explored decision-theoretic models where agents derive utility from beliefs. In one strand of the literature demand for self-esteem management arises from a need to compensate for a lack of self-control (Carrillo and Mariotti 2000, Benabou and Tirole 2002). In the other strand, agents derive direct utility from their beliefs (Akerlof and Dickens 1982, Caplin and Leahy 2001, Koszegi 2006, Brunnermeier and Parker 2005). These papers review compelling evidence that well-being is directly affected by beliefs about the future. For example, knowing that one will eat dinner with old friends tomorrow induces pleasant feelings of anticipation, while knowing that one will undergo chemotherapy tomorrow excites dread – dread strong enough to induce vomiting in some subjects.

Our model builds on Brunnermeier and Parker’s (2005) model of “anticipatory utility”. Their approach provides an unified account of both the benefits and the costs of self-esteem management: biasing one’s beliefs up raises anticipatory utility but also increases the risk of making a mistaken over-confident choice. Moreover, since the “beliefs” component of utility is derived from rational expectations about future payoffs, its structure is not a free parameter but rather is pinned down by the agent’s environment.

**Investment cost.** We assume a uniform cost distribution over $[\underline{c}, 1]$ with $\mu < \underline{c}$. Our results do not depend on the uniformity of the distribution which we impose for analytical tractability.
Assuming \( \mu < \mu \) introduces a new effect: biasing one’s belief now always has a cost, including the low state of the world. This partially mitigates the incentive to bias one’s belief and the extent of the net bias depends on details of the distribution. We focus on the case \( \mu < \mu \) because it maximizes the demand for learning as the uninformed agent would never invest.

3 Experimental Design and Methodology

Our experiment consisted of four stages which are summarized in Figure 5 and explained in detail below. During the quiz stage, a subject completed an online IQ test. During the feedback stage we provided each subject with four binary signals which indicated whether the subject was among the top half of performers. We measured each subject’s belief about being among the top half of performers six times – before the IQ quiz, after the IQ quiz and after each signal. In the information purchasing stage we gave subjects the opportunity to purchase precise information about whether her performance put her in the top half of all performers. A sub-sample of subjects were invited one month later for a follow-up which repeated the feedback stage but with reference to the performance of a robot rather than to their own performance.

3.1 Quiz Stage

Subjects had four minutes to answer as many as possible out of 30 questions. Since the experiment was web-based and different subjects took the test at different times we randomly assigned each subject to one of 9 different versions of the IQ test. Subjects were informed that their performance would be compared to the performance of all other students taking the same
test version. We chose an IQ test because we wanted a task that subjects care strongly about. The tests consisted of standard logic 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 subject’s final score was the number of correct answers minus the number of incorrect answers. Earnings for the quiz were the score multiplied by $0.25. During the same period we conducted an unrelated experiment on social learning and the combined earnings of all parts of all experiments were transferred to subjects’ university debit cards at the end of the study. Since we did not itemize earnings and earnings were variable (and even differed across IQ tests) it would have been very difficult for subjects to infer performance from earnings.

**Types.** Subjects with IQ scores above the median for their particular IQ quiz correspond to high types in our model, those with scores below the median to low types. Because types are binary a subject’s belief about her type 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 much of the literature where only several moments of more complicated belief distributions are elicited.\(^\text{11}\)

### 3.2 Feedback Stage

**Signal Accuracy.** Signals were independent and correct with probability 75 percent: if a subject was among the top half of performers she would get a “Top” signal with probability \(p = 0.75\) and a “Bottom” signal with probability \(1 - p\). If a subject was among the bottom half of performers she would get a Top signal with probability \(q = 0.25\) and a Bottom signal with probability \(1 - q\). The informativeness of Top and Bottom signals was therefore \(\lambda_H = \ln(3)\) and \(\lambda_L = -\ln(3)\), respectively. To explain the accuracy of signals over the web, subjects 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.** We used a novel crossover mechanism each time we elicited beliefs. Subjects were presented with two options,

\(^{11}\)For example, Niederle and Vesterlund (2007) elicit the mode of subjects beliefs about their rank in groups of 4.
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. To present this mechanism in a simple narrative form, we told subjects that they were paired with a “robot” who had a fixed but unknown probability $x$ between 0 and 100% of scoring among the top half of subjects. Subjects 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 to subjects that they would maximize their probability of earning the $3 by choosing as the threshold their own subjective probability of being in the top half. Subjects were told at the outset that we would elicit their beliefs several times but would implement only one choice at random for payment.

To the best of our knowledge ours is the first paper to implement the crossover mechanism in an experiment. The crossover mechanism has two main advantages over the quadratic scoring rule commonly used in experimental papers. First, quadratic scoring is truth-inducing only for risk-neutral subjects; the crossover mechanism is strictly incentive-compatible provided only that subjects’ preferences are monotonic in the sense that among lotteries that pay $3 with probability $q$ and $0$ with probability $1 - q$ they strictly prefer those with higher $q$. This property holds for all von-Neumann-Morgenstern preferences as well as for many non-standard preferences such as Prospect theory.

A second advantage is that the crossover mechanism does not generate perverse incentives to “hedge” performance on the quiz. To see the issue, consider the incentives facing a subject who has predicted that she will score in the top half with probability $\hat{\mu}$. Under a quadratic scoring rule she will earn a piece of $0.25 per point she scores and lose an amount proportional to $(1(S \geq \overline{S}) - \hat{\mu})^2$, where $S$ is her score and $\overline{S}$ the median score. If she believes the latter to be distributed $F$ then her total payoff is

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

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

$$ S \times 0.25 + 3.00 \times \hat{\mu} \times \int_{\overline{S}} 1(S \geq \overline{S}) dF(S) $$

12 After running our experiment we became aware that the same mechanism was also independently discovered by Allen (1987) and Grether (1992), and has since been proposed by Karni (2009).

13 See Offerman, Sonnemans, Van de Kuilen and Wakker (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 subjects, but unpaid subjects tend to report inaccurate and incoherent beliefs (Grether 1992).
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 as best she can.

3.3 Information Purchasing Stage

In the final stage of the experiment we elicited subjects’ demand for noiseless feedback on their relative performance. We asked subjects to state their willingness to pay for the following bundles: receiving $2$, receiving $2$ and receiving feedback through a private email, or receiving $2$ and receiving feedback on a web page visible to all study participants. We offered two variants of the latter two bundles, one in which subjects learned whether they scored in the top half or not and another in which they learned their exact quantile in the score distribution. In total subjects thus bid for five bundles. We bounded responses between $0.00$ and $4.00$.

One of the choices was randomly selected and subjects purchased the corresponding bundle if and only if their reservation price exceeded a randomly generated price. This design is a standard application of BDM, with the twist that we measure information values by netting out subjects’ valuations for $2$ alone from their other valuations. This addresses the concern that subjects may under-bid for objective-value prizes.

3.4 Follow-up Stage

We invited a random sub-sample of subjects through email to a follow-up experiment one month later. Subjects were told they had been paired with a robot who had a probability $\theta$ of being of high type. We then repeated the feedback stage of the experiment except that this time subjects received signals of the robot’s ability and we tracked their beliefs about the robot being a high type.

The purpose of this follow-up was to compare subjects’ processing of information about a robot’s ability as opposed to their own ability. To make this within-subject treatment 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, $\theta$, to the multiple of 5% closest to the subject’s post-IQ quiz confidence. For example, if the subject had reported a confidence level of 63% after the quiz we would pair the subject with a robot that was a high type with probability $\theta = 65\%$. We then randomly picked a high or low type robot for each subject with probability $\theta$. If the type of the robot matched the subject’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.
4 Data

4.1 Subject 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 home page on www.facebook.com, a popular social networking site. 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. This final 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.

An important concern with an online experiment is whether subjects understood and were willing to follow the instructions. To account for this, our software required subjects to make an active choice each time they submitted a belief – they were free to report beliefs that are clearly inconsistent with Bayesian updating such as updates in the wrong direction and neutral updates (reporting the same belief as in the previous round). After each of the 4 signals, a stable proportion of about 36 percent of subjects reported the same belief as in the previous round. About 16% of subjects did not change their beliefs at all during all four rounds of feedback. In contrast, the share of subjects who updated in the wrong direction declined over time (12.67%, 8.70%, 7.66% and 6.62%) and most subjects made at most one mistake.

For most of our analysis we use a restricted sample of subjects who (1) made no updates in the wrong direction and (2) revised their beliefs at least once. These restrictions exclude 25% and 13% of our sample, respectively, and leave us with 341 women and 314 men. We view this exclusion as a conservative way to exclude subjects who misunderstood or ignored the instructions. Our main conclusions hold on the full sample as well, however, and we also provide those estimates as robustness checks where appropriate.

We invited 120 subjects to participate in the follow-up stage one month later and 78 completed this final stage of the experiment. The pattern of wrong and neutral moves was similar to the first stage of the experiment. Slightly fewer subject made neutral updates (28% of all updates) and 10% always made neutral updates. Slightly more subjects made wrong

\[14\text{In November 2004 more than 90\% of students were members of the site and at least 60\% of members logged into the site daily.}\]

\[15\text{The exact proportions were 35.82\%, 38.85\%, 37.43\% and 35.92\%.}\]

\[16\text{Overall, 18.82\% of subjects made only one mistake, 6.43\% made two mistake, 1.51\% made 3 mistakes and 0.38\% made 4 mistakes.}\]
Empirical cumulative distributions of subjects’ beliefs directly following the quiz and after four rounds of noisy feedback.

updates (21.79% made one mistake, 10.26% made two mistakes, 5.13% made three mistakes and 2.56 made 4 mistakes). The restricted sample for the follow-up has 39 subjects.

4.2 Summary Statistics

Quiz scores. The mean score of the 656 subjects 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) is approximately normal with a handful of outliers who appear to have guessed randomly. The most questions answered by a subject was 29, so the 30-question limit did not induce bunching at the top of the distribution. Appendix Table C.1 provides further descriptive statistics broken down by gender and by quiz type. An important observation is that the 9 versions of the quiz varied substantially in difficulty, with
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 subjects were Bayesian are provided for comparison. T-bars indicate 95% confidence intervals.

Mean scores on the easiest version (#6) five times higher than on the hardest version (#5). Subjects 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. (Moore and Healy (2008) document a similar pattern.) We will exploit this variation below, using quiz assignment as an instrument for beliefs.

Belief updating. Figure 6 plots the empirical cumulative distribution function of subjects’ beliefs 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).

Our design with only two states (Top and Bottom half of the distribution) allows us to
Mean absolute belief revisions by decile of prior belief in being of type equal to the signal received. For example, a subject with prior belief $\mu = 0.8$ of being in the top half who received a signal $T$ and a subject with prior belief $\mu = 0.2$ who receive a signal $B$ are both plotted at $x = 80\%$. T-bars indicate 95\% confidence intervals.

easily compare the belief updates of subjects to Bayesian updates. Figure 7 shows the mean belief revision in response to a Top and Bottom signal by decile of prior belief in being a Top type for each of the four observations of the 656 subjects. Our first observation is that subjects are conservative and update much less than the Bayesian benchmark would predict. Second, despite the fact that signals are symmetric, subjects update asymmetrically. Figure 8 compares subjects whose prior belief was $\mu$ and who received positive feedback with subjects who prior belief was $1 - \mu$ and who receive negative feedback. According to Bayes rule, the magnitude of the belief change in these situations should be identical, but in fact subjects tend to respond more strongly to positive feedback. We will study both of these phenomena using a regression approach in the next section.
### Table 1: Implied Valuations for Information: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>$P(v &lt; 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimation Sample</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information (Coarse)</td>
<td>650</td>
<td>16.5</td>
<td>47.8</td>
<td>0.09</td>
</tr>
<tr>
<td>Information (Precise)</td>
<td>650</td>
<td>40.0</td>
<td>78.3</td>
<td>0.09</td>
</tr>
<tr>
<td>Publicity (Coarse)</td>
<td>651</td>
<td>-52.3</td>
<td>73.0</td>
<td>0.66</td>
</tr>
<tr>
<td>Publicity (Precise)</td>
<td>651</td>
<td>-71.1</td>
<td>88.0</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information (Coarse)</td>
<td>338</td>
<td>16.4</td>
<td>49.8</td>
<td>0.11</td>
</tr>
<tr>
<td>Information (Precise)</td>
<td>338</td>
<td>38.7</td>
<td>82.0</td>
<td>0.11</td>
</tr>
<tr>
<td>Publicity (Coarse)</td>
<td>339</td>
<td>-57.4</td>
<td>74.3</td>
<td>0.72</td>
</tr>
<tr>
<td>Publicity (Precise)</td>
<td>339</td>
<td>-77.1</td>
<td>89.6</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information (Coarse)</td>
<td>312</td>
<td>16.7</td>
<td>45.5</td>
<td>0.07</td>
</tr>
<tr>
<td>Information (Precise)</td>
<td>312</td>
<td>41.5</td>
<td>74.1</td>
<td>0.06</td>
</tr>
<tr>
<td>Publicity (Coarse)</td>
<td>312</td>
<td>-46.7</td>
<td>71.2</td>
<td>0.60</td>
</tr>
<tr>
<td>Publicity (Precise)</td>
<td>312</td>
<td>-64.5</td>
<td>85.7</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Values for coarse information (learning whether you were in the top or bottom half) and precise information (learning your exact percentile rank) are the differences between subjects' bids for $2 and their bids for the bundle of $2 and learning that information via email. Values for publicity are the differences between bids for obtaining information by email and bids for additionally having it made public online. Values are in cents. The final column reports the fraction of observations with strictly negative valuations. There are fewer than 656 observations because 6 (5) subjects did not provide valuations for coarse (precise) information.

**Value of information.** We impute the value for the various information packages. For example, a subjects valuation for coarse information on their performance - whether or not they were in the top half - is defined as their bid for $2 and learning whether they scored in the top half minus their bid for $2, all in cents. Her valuation for publicity of that coarse information is her bid for receiving information publicly minus her bid for receiving it privately. Subjects’ mean value for coarse information of 16.5 (s.d. 47.8), with 9% of subjects reporting a negative value. The value of publicity for coarse information is -52.3 (s.d. 73.0) with 66% of subjects reporting negative values. Subjects could also receive precise information, their precise quantile. The value for private precise information exceeded that of coarse information: the mean value was 40.0 (s.d. 78.3) with 9% of subjects reporting a negative value. Once more publicity was viewed as much less desirable, with a mean value of -71.1 (s.d. 88.0) and 71% of subjects reporting a negative value. We will return to these observations in Section 6.
5 Information Processing

In this section we analyze belief updating in the feedback and follow-up stages, comparing our model’s predictions to the Bayesian benchmark.

5.1 Empirical Specification

Our empirical strategy mirrors the theory section, expressing information processing in terms of the logistic function. For a (possibly biased) Bayesian,

\[
\logit(\hat{\mu}_t) = \logit(\hat{\mu}_{t-1}) + I(s_t = H) \cdot \hat{\lambda}_H + I(s_t = L) \cdot \hat{\lambda}_L
\]

(19)

This motivates the following linear empirical specification:

\[
\logit(\hat{\mu}_t^i) = \delta \cdot \logit(\hat{\mu}_{t-1}^i) + \beta_H \cdot I(s_{it} = H) \lambda_H + \beta_L \cdot I(s_{it} = L) \lambda_L + \epsilon_{it}
\]

(20)

In our experiment, we have \(\lambda_H = -\lambda_L = \log(3)\) and the error term \(\epsilon_{it}\) captures unsystematic errors that subject \(i\) when updating her belief at time \(t\). Note that we do not have to include a constant in this regression because \(I(s_{it} = H) + I(s_{it} = L) = 1\). The coefficient \(\delta\) captures the persistence of prior information; our model predicts \(\delta = 1\) for both biased and perfect Bayesians. The coefficients \(\beta_H\) and \(\beta_L\) capture responsiveness to positive and negative information and allow us to distinguish perfect and biased Bayesians. A perfect Bayesian is fully responsive to positive and negative information (\(\beta_H = \beta_L = 1\)). In contrast a biased Bayesian is conservative – less responsive to new information overall (\(\beta_H, \beta_L < 1\)) – and asymmetric – more responsive to positive than negative information (\(\beta_H > \beta_L\)).

Identifying equation (20) is non-trivial because we include lagged logit-beliefs (i.e. priors) as a dependent variable. If there is unobserved heterogeneity in subjects’ responsiveness to information, \(\beta_L\) and \(\beta_H\), then OLS estimation may yield upwardly biased estimates of \(\delta\) 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 subjects make mistakes or are imprecise in recording their beliefs.

\[\text{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 (e.g. 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 \(\delta\) is close to one, as our model predicts for both perfect and biased Bayesians.}\]
To address these issues we exploit the fact that subjects’ 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 subject $i$ we calculate the average quiz score of subjects other than $i$ who took the same quiz variant to obtain a measure of the quiz difficulty level that is not correlated with subject $i$’s own ability but highly correlated with subjects beliefs. We will report both OLS and IV estimates of Equation 20.

5.2 Results from Feedback Stage

Table 2 presents round-by-round and pooled estimates of Equation 20. Estimates in Panel A are via OLS and those in Panel B 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 2002).

Result 1 (Persistence) Subjects weigh prior information similarly to Bayesian updaters.

Our model implies a coefficient $\delta = 1$ on prior logit-beliefs for both perfect and biased Bayesians. OLS estimates for the early rounds of belief updating put it close to but significantly less than unity. It climbs by round, however, until by Round 4 we cannot reject the hypothesis that it equals one ($p = 0.57$). These estimates may be biased upward by heterogeneity in the responsiveness coefficients, $\beta_iL$ and $\beta_iH$, or may be biased downwards if subjects report beliefs with noise. The IV estimates suggest that the latter bias is more important: the pooled point estimate is larger 0.963 and none of the estimates are significantly different from unity. All told, we find strong evidence that information persists once it has been incorporated into agents’ beliefs.

Result 2 (Conservatism) Subjects respond less to both positive and negative information than a perfect Bayesian.

Figure 7 suggests that our subjects respond less to new information than a perfect Bayesian. This observation is reflected in the regressions. Our OLS estimates of $\beta_H$ and $\beta_L$, 0.370 and 0.302, are substantially and significantly less than unity. Round-by-round estimates do not follow any obvious trend – therefore, this observation does not seem to be a mere cognitive error that can be reduced through practice. The IV and OLS estimates are almost identical, suggesting there is little bias through correlation with lagged prior beliefs.

---

18 We must restrict ourselves to observations for which both posteriors and priors were in $(0, 1)$ so that $\text{logit}(\hat{\mu}_t^i)$ and $\text{logit}(\hat{\mu}_{t-1}^i)$ are defined; to balance the panel we further restrict the sample to subjects $i$ for whom this holds for all rounds $t$. Results from the ragged panel, which includes another 101 observations, are essentially identical.
Table 2: Conservative and Asymmetric Belief Updating

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
<th>All Rounds</th>
<th>Unrestricted</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta)</td>
<td>0.814</td>
<td>0.925</td>
<td>0.942</td>
<td>0.987</td>
<td>0.924</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td>(0.030)**</td>
<td>(0.015)**</td>
<td>(0.023)**</td>
<td>(0.022)**</td>
<td>(0.011)**</td>
<td>(0.014)**</td>
</tr>
<tr>
<td>(\beta_H)</td>
<td>0.374</td>
<td>0.295</td>
<td>0.334</td>
<td>0.438</td>
<td>0.370</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>(0.019)**</td>
<td>(0.017)**</td>
<td>(0.021)**</td>
<td>(0.030)**</td>
<td>(0.013)**</td>
<td>(0.013)**</td>
</tr>
<tr>
<td>(\beta_L)</td>
<td>0.295</td>
<td>0.274</td>
<td>0.303</td>
<td>0.347</td>
<td>0.302</td>
<td>0.211</td>
</tr>
<tr>
<td></td>
<td>(0.025)**</td>
<td>(0.020)**</td>
<td>(0.022)**</td>
<td>(0.024)**</td>
<td>(0.012)**</td>
<td>(0.011)**</td>
</tr>
<tr>
<td>(P(\beta_H = 1))</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(P(\beta_L = 1))</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(P(\beta_H = \beta_L))</td>
<td>0.009</td>
<td>0.408</td>
<td>0.305</td>
<td>0.017</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>N</td>
<td>612</td>
<td>612</td>
<td>612</td>
<td>612</td>
<td>2448</td>
<td>3996</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.803</td>
<td>0.890</td>
<td>0.875</td>
<td>0.859</td>
<td>0.854</td>
<td>0.798</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(\beta_H)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(\beta_L)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(P(\beta_H = 1))</td>
</tr>
<tr>
<td>(P(\beta_L = 1))</td>
</tr>
<tr>
<td>(P(\beta_H = \beta_L))</td>
</tr>
<tr>
<td>First Stage F-statistic</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>(R^2)</td>
</tr>
</tbody>
</table>

Notes:
1. Each column in each panel is a regression. The outcome in all regressions is the log posterior odds ratio. \(\delta\) is the coefficient on the log prior odds ratio; \(\beta_H\) and \(\beta_L\) 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 subjects whose beliefs were always within (0,1). Columns 1-5 further restrict to subjects who updated their beliefs at least once and never in the wrong direction; Column 6 includes subjects 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 subjects 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: \(*p < 0.10\), \(**p < 0.05\), \(***p < 0.01\).
Result 3 (Asymmetry) Controlling for prior beliefs subjects respond more to positive signals than to negative ones.

The regressions also confirm that subjects respond differently to positive and negative information as suggested in figure 8. To quantify asymmetry we compare estimates of $\beta_H$ and $\beta_L$, the responsiveness to positive and negative signals. The difference $\beta_H - \beta_L$ is consistently positive across all rounds and significantly different from zero in the first round, fourth round, and for the pooled specification. While estimates of this difference in rounds 2 and 3 are not significantly different from zero, we cannot reject the hypothesis that the estimates are equal across all four rounds ($p = 0.32$). The IV estimates are somewhat more variable but are again uniformly positive and significantly so in rounds 1 and 4 and in the pooled specification. The size of the difference is substantial, implying that the effect of receiving both a positive and a negative signal (i.e., no information) is 26% as large as the effect of receiving only a positive signal. As an alternative non-parametric test we can study the net change in beliefs among the 224 subjects who received two positive and two negative signals. These subjects should have ended with the same beliefs as they began; instead their beliefs increased by an average of 4.8 points ($p < 0.001$).

A key benefit of our empirical design is that it not only rejects the Bayesain model but shows us in exactly which ways it fails. If instead we simply regress subjects logit-beliefs on those predicted by Bayes rule we estimate a correlation of 0.57, which lets us reject the Bayesian null but does not disentangle persistence, conservatism, or asymmetry. On the other hand, it is interesting to summarize the extent to which subjects deviate from Bayesian updating by comparing their payoffs $\pi_{\text{actual}}$ to those they would have earned if they updated using Bayes rule ($\pi_{\text{Bayes}}$) or if they reported uniformly random posteriors ($\pi_{\text{random}}$). The ratio $\frac{\pi_{\text{actual}} - \pi_{\text{random}}}{\pi_{\text{Bayes}} - \pi_{\text{random}}}$ is 0.64, implying that non-Bayesian updating behavior cost subjects 36% of the potential gains from processing information within this experiment.

5.3 Results from Follow-up Stage

While our model of self-confidence management explains both conservatism and asymmetry, there are other interpretations unrelated to ego that might explain some of our results. For example, conservatism might arise if subjects are perfect Bayesians who simply misinterpret the informativeness of signals and believe that the signal is only correct with 60% probability instead of 75%. Subjects might underweight our signals because they are used to encountering weaker ones in everyday life. To distinguish our model from these interpretations we analyze the results of the follow-up experiment, in which a random subset of subjects performed an updating task that was formally identical to the one in the original experiment but which
Table 3: Belief Updating: Own v.s. Robot Performance

<table>
<thead>
<tr>
<th>Regressor</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_H$</td>
<td>0.426</td>
<td>0.349</td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td>(0.087)**</td>
<td>(0.066)**</td>
<td>(0.043)**</td>
</tr>
<tr>
<td>$\beta_L$</td>
<td>0.330</td>
<td>0.241</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>(0.050)**</td>
<td>(0.042)**</td>
<td>(0.033)**</td>
</tr>
<tr>
<td>$\gamma_H$</td>
<td>0.362</td>
<td>0.227</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>(0.155)**</td>
<td>(0.116)*</td>
<td>(0.081)</td>
</tr>
<tr>
<td>$\gamma_L$</td>
<td>0.356</td>
<td>0.236</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.120)**</td>
<td>(0.085)**</td>
<td>(0.089)</td>
</tr>
<tr>
<td>$P(\beta_H + \gamma_H = 1)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$P(\beta_L + \gamma_L = 1)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.088</td>
</tr>
<tr>
<td>$P(\beta_H + \gamma_H = \beta_L + \gamma_L)$</td>
<td>0.454</td>
<td>0.316</td>
<td>0.030</td>
</tr>
<tr>
<td>N</td>
<td>160</td>
<td>248</td>
<td>480</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.567</td>
<td>0.434</td>
<td>0.114</td>
</tr>
</tbody>
</table>

Notes:

1. Each column is separate regression. The outcome in all regressions is the change in the log belief ratio. $\beta_H$ and $\beta_L$ are the estimated effects of the log likelihood ratio for positive and negative signals, respectively. $\gamma_H$ and $\gamma_L$ are the differential response attributable to obtaining a signal about the performance of a robot as opposed to about one’s own performance.

2. Estimation samples are restricted to subjects who participated in the follow-up experiment and observed the same sequence of signals as in the main experiment. Column I includes only subjects who updated at least once in the correct direction and never in the wrong direction in both experiments. Column II adds subjects who never updated their beliefs. Column III includes all subjects.

3. Robust standard errors clustered by individual reported in parentheses. Statistical significance is denoted as: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$
dealt with the ability of a robot rather than their own ability. For these subjects we pool the updating data from both experiments and estimate

\[
\text{logit}(\hat{\mu}_t^e) - \text{logit}(\hat{\mu}_{t-1}^e) = \beta_H \cdot I(s_{it} = H)\lambda_H + \beta_L \cdot I(s_{it} = L)\lambda_L + \\
+ \gamma_H \cdot 1(e = \text{Robot}) \cdot I(s_{it} = H)\lambda_H + \gamma_L \cdot 1(e = \text{Robot}) \cdot I(s_{it} = L)\lambda_L + \epsilon_t
\]

(21)

Here \( e \) indexes experiments (Ego or Robot), so that the interaction coefficients \( \gamma_H \) and \( \gamma_L \) tell us whether subjects’ process identical information differently across both treatments. Note that as we cannot reject \( \delta = 1 \) (Table 2) we impose that restriction here for brevity.

**Result 4** Both conservatism and asymmetry are reduced, the former significantly, when the same subjects with the same initial priors observe the same flow of information about a robot’s performance rather than their own performance.

Table 3 reports results. The baseline coefficients \( \beta_H \) and \( \beta_L \) are similar to their estimated values for the larger sample (see Table 2), suggesting that participation in the followup was not selective on updating traits. The interaction coefficients are both positive and significant – they imply that subjects 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 \( \beta_H + \gamma_H = 1 \) (\( p = 0.13 \)), though we can still reject \( \beta_L + \gamma_L = 1 \) (\( p = 0.004 \)). While conservatism does not entirely vanish it is clearly much weaker. This provides support for our model where conservatism is driven by anticipatory ego utility. Also consistent with this view is the fact that subjects are less asymmetric in relative terms when they update about robot performance (Theorem 1), since \( \frac{\beta_H}{\beta_L} > \frac{\beta_H + \gamma_H}{\beta_L + \gamma_L} \). We cannot reject the hypothesis that they update symmetrically about robot performance such that \( \beta_H + \gamma_H = \beta_L + \gamma_L \) (\( p = 0.45 \)).

5.4 Discussion

We now discuss how our results relate to the existing experimental literature on information processing and self-confidence.

**Information Retention.** The mainstream approach to modeling belief evolution – which includes both the Bayesian model and our own “biased Bayesian” framework – assumes that information once incorporated into beliefs is persistent. Our results provide empirical support – the first we are aware of – for this basic assumption. They do not, of course, rule out information loss over longer periods. In this vein Mullainathan (2002), Wilson (2003), and Benabou and Tirole (2002) examine the implications of imperfect memory.

**Attribution bias.** A large literature in psychology has argued for the existence of self-serving “attribution biases,” or tendencies to take credit for good outcomes and deny blame
for bad ones. Critics have pointed out, however, that most of these studies “seem readily interpreted in information-processing terms” (Miller and Ross 1975, p. 224).

One common problem is that the data-generating process is not clearly defined, so that the evidence can be rationalized using Bayes rule and appropriate subjective beliefs (Wetzel 1982). For example, in one prototypical experimental paradigm subjects were asked to teach a student and then asked in an open-ended format whether the student’s performance was attributable to their teaching or to other factors. Researchers took as evidence for a self-serving bias the fact that subjects often attribute poor performances to lack of effort on the part of the student while taking credit for good performances. But this is obviously consistent with rational inference if one believes that student effort and teacher ability are complementary in producing success.

Another common issue is that key outcome variables are not objectively defined or elicited with incentives for honesty. For example, Wolosin, Sherman and Till (1973) had subjects 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...” (1387).

In contrast to these studies, we (1) clearly define the probabilistic event (scoring in the top half) and outcome variables (subjective beliefs about the probability of that event) of interest, and (2) explicitly inform subjects about the conditional likelihood of observing different signals. The lack of ambiguity makes our test for asymmetry unconfounded and also stringent, since it may well be precisely in the interpretation of ambiguous concepts that agents are most free to be self-serving. We find asymmetry nevertheless.

**Overconfidence.** Over time asymmetric updating leads to *overconfidence*, in the sense that individuals will over-estimate their probability of succeeding at a task compared to a Bayesian’s forecast. We emphasize this definition to contrast it with others frequently used in the literature. Findings that more than $x\%$ of a population believe that they are in the top $x\%$ in terms of some desirable trait are commonly taken as evidence of irrational overconfidence, but Zábojník (2004), Van den Steen (2004), Santos-Pinto and Sobel (2005), and Benoit and Dubra (2008) have all illustrated how such results can obtain under strictly Bayesian information processing. Our definition and result are not subject to these critiques.

**Conservatism and Bayes rule.** Psychologists have also tested Bayes rule as a positive model of human information-processing in ego-neutral settings. A prototypical experiment involves showing subjects two urns containing 50% and 75% red balls respectively, then showing them a sample of balls drawn from one of the two urns and asking them to predict which urn was
used. Unsurprisingly these studies do not find asymmetry (indeed it is unclear how one would define it when ego is not at stake). Studies during the 1960s did find conservatism, but this view was upset by Kahneman and Tversky’s (1973) discovery of the “base rate fallacy,” seen as “the antithesis of conservativism” (Fischhoff and Beyth-Marom 1983, 248-249). Recently Massey and Wu (2005) have generated both conservative and anti-conservative updating within a single experiment: their subjects underweight signals with high likelihood ratios but overweight signals with low likelihood ratios. In the light of this literature it is important that we find significantly more conservatism when subjects update about their own performance as opposed to a robot’s performance, holding constant the data generating process. This supports the interpretation that conservatism is a motivated and not merely a cognitive bias.

**Confirmatory bias.** Asymmetry is not obviously more pronounced among subjects with more optimistic prior (see figure 5). This is not consistent with at least simple interpretations of confirmatory bias (Rabin and Schrag 1999). However, our results do mechanically imply a steady-state relationship similar to confirmatory bias: more asymmetric individuals will tend both to have higher beliefs and to respond more to positive information.

### 6 Demand for Information

The standard economic model of learning predicts that agents always place a weakly positive value on information. This is because the best action to take after receiving information cannot do worse on average than the action one would have taken without it. This need not hold in our model, however (Theorem 2), because hard information tends to destroy the anticipatory utility built up through asymmetric updating.

To distinguish these views we define subject $i$’s revealed value of information $v_i$ as the difference between their willingness to pay for “$2 and learning their actual performance” and their willingness to pay for $2 alone. As discussed earlier, taking this differences removes bias due to misunderstanding the dominant strategy in the “bid for $2” decision problem.\footnote{89\% of our subjects bid less than $2, and 80\% bid less than $1.99.}

The first two rows of Table 1 provide summary statistics for information valuations: learning via email whether or not one scored in the top half (coarse), or learning one’s exact rank (precise). (The second two rows summarize how these valuations change when the information is delivered publicly on a web page that other participants can view in addition to privately via email.) Mean valuations are positive and more precise information is valued more on average. A substantial fraction, however, place a strictly negative valuation on coarse information: 9\% of subjects overall, 11\% of women and 7\% of men. Almost identical fractions place strictly
negative valuations on precise information about their rank.\footnote{As an interesting contrast, Eliaz and Schotter (forthcoming) document that subjects are willing to pay positive amounts for information (unrelated to ego) even when there is no probability it can improve their decision-making.}

**Result 5 (Information Aversion)** A substantial fraction of subjects are willing to pay to avoid learning their type.

A potential concern about this result is that it could be an artefact attributable to noise in subjects’ recorded valuations. One piece of evidence that suggests this is not the case is the high correlation ($\rho = 0.77$) between having a negative valuation for coarse information and a negative valuation for precise information, which suggests a systematic pattern. Building on this intuition, Appendix B shows that under the weak structural assumption that reporting errors are independently normally distributed the recording error interpretation is rejected by the second moments of the bid data.

**Result 6** More confident subjects are less information-averse.

In addition to predicting information-aversion Theorem 2 implies that it should be less common among more confident agents. To test this implication we regress an indicator $I(v_i \geq 0)$ on subjects’ logit posterior belief after all four rounds of updating, which is when they bid for information. Columns I-III of Table 4 show that, as predicted, subjects with higher posterior beliefs are significantly more likely to have (weakly) positive information values. The point estimate is slightly larger and remains strongly significant when we control for ability ability (Column II) and gender and age (Column III). Of course, there could be some 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 subjects randomly assigned to the same quiz type remains a valid instrument for beliefs as in Section 5 above. In addition, once we control for whether or not the subject scored in the top half, the number of positive signals they received during the updating stage is also a valid instrument, since conditional on type the signals were generated randomly. Estimates using these instruments are similar to the OLS estimates, slightly larger, and though less precise are still significant at the 10% level.

### 7 Gender Differences

Gender differences related to self-confidence have been documented in a variety of settings.\footnote{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...} It is therefore natural to ask whether men and women process and acquire confidence-relevant...
Table 4: Confidence and Positive Information Value

<table>
<thead>
<tr>
<th>Regressor</th>
<th>OLS I</th>
<th>OLS II</th>
<th>OLS III</th>
<th>OLS IV</th>
<th>OLS V</th>
<th>IV I</th>
<th>IV II</th>
<th>IV III</th>
<th>IV IV</th>
<th>IV V</th>
</tr>
</thead>
<tbody>
<tr>
<td>logit(μ)</td>
<td>0.017</td>
<td>0.023</td>
<td>0.023</td>
<td>0.027</td>
<td>0.027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.007)**</td>
<td>(0.009)**</td>
<td>(0.009)**</td>
<td>(0.016)*</td>
<td>(0.017)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top Half</td>
<td>-0.033</td>
<td>-0.035</td>
<td>-0.038</td>
<td>-0.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.034)</td>
<td>(0.034)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.029</td>
<td>0.027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.023)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YOG</td>
<td>0.018</td>
<td>0.018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-Stage F-Statistic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>118.48</td>
<td>113.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>609</td>
<td>609</td>
<td>609</td>
<td>609</td>
<td>609</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.007</td>
<td>0.010</td>
<td>0.016</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Each column is a separate regression. Estimation is via OLS 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 subject’s valuation for information was positive; the mean of this variable is 0.91. “Top Half” is an indicator equal to one if the subject scored above the median on his/her quiz type; “YOG” is the subject’s year of graduation. Heteroskedasticity-robust standard errors in parenthesis. Statistical significance is denoted as: *p < 0.10, **p < 0.05, ***p < 0.01.

information differently and, if so, whether these differences are well-explained by our theory of self-confidence maintenance. This section examines these issues. We begin with information processing. Table 5 reports estimates of Equation 20 differentiated by gender (and also by ability) and estimated using both OLS and instrumental variables. Men are substantially less conservative than women, reacting significantly more to both positive and negative feedback than women and 21% more to feedback on average (23% in when estimated by GMM). Estimated changes in relative asymmetry are less stable; OLS and IV point estimates of $\beta_{H} + \beta_{Male} - \beta_{L}$ 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.

Turning to demand for feedback, men and women place similar average valuations on information; the means reported in Table 1 are not statistically different from each other.

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.

Consistent with prior work we find that men are significantly more confident than women. The mean difference in confidence prior to taking the quiz was 6.7% ($p < 0.001$). Some of this could reflect differences in actual ability: men scored 7.9 on average while women scored 6.9, and this difference is highly significant ($p < 0.001$). Even when we restrict ourselves to variation within groups of subjects who took the same version of the quiz and received the same score, however, we find that men are 5.0% more confident on average and this difference remain highly significant ($p < 0.001$).
## Table 5: Heterogeneity in Updating Along Observable Dimensions

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Gender</th>
<th>Ability</th>
<th>Gender</th>
<th>Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td>0.925</td>
<td>0.918</td>
<td>0.988</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>(0.015)***</td>
<td>(0.015)***</td>
<td>(0.103)***</td>
<td>(0.075)***</td>
</tr>
<tr>
<td>$\delta^{Male}$</td>
<td>-0.007</td>
<td>0.010</td>
<td>-0.047</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td></td>
<td>(0.120)</td>
<td></td>
</tr>
<tr>
<td>$\delta^{Able}$</td>
<td>0.010</td>
<td>-0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta^H$</td>
<td>0.331</td>
<td>0.381</td>
<td>0.344</td>
<td>0.407</td>
</tr>
<tr>
<td></td>
<td>(0.017)***</td>
<td>(0.026)***</td>
<td>(0.031)***</td>
<td>(0.050)***</td>
</tr>
<tr>
<td>$\beta^L$</td>
<td>0.280</td>
<td>0.317</td>
<td>0.258</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>(0.015)***</td>
<td>(0.016)***</td>
<td>(0.040)***</td>
<td>(0.034)***</td>
</tr>
<tr>
<td>$\beta^{Male}^H$</td>
<td>0.080</td>
<td>0.063</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.027)***</td>
<td></td>
<td>(0.038)*</td>
<td></td>
</tr>
<tr>
<td>$\beta^{Male}^L$</td>
<td>0.052</td>
<td>0.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.026)**</td>
<td></td>
<td>(0.044)*</td>
<td></td>
</tr>
<tr>
<td>$\beta^{Able}^H$</td>
<td>-0.017</td>
<td></td>
<td>-0.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td></td>
<td>(0.054)</td>
<td></td>
</tr>
<tr>
<td>$\beta^{Able}^L$</td>
<td>-0.041</td>
<td></td>
<td>-0.011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td></td>
<td>(0.049)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>2448</td>
<td>2448</td>
<td>2448</td>
<td>2448</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.855</td>
<td>0.854</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Notes:

1. Each column is a separate regression; the outcome variable in all regressions is the log posterior odds ratio. The coefficients $\delta$, $\beta^H$, and $\beta^L$ are the estimated effects of the log prior odds ratio, log likelihood ratio for positive signals, and log likelihood ratio for negative signals, respectively. Superscripted coefficients report the differential effects of these variables for men and for the “able”, i.e. those who scored in the top half.

2. Columns 1 and 2 report OLS estimates; Column 3 (4) reports GMM results using the average score achieved by other subjects randomly assigned to the same quiz type the interaction of this variable with male (ability) as instruments.

3. Robust standard errors clustered by individual (OLS) or HAC standard errors (GMM) in parenthesis. Statistical significance is denoted as: $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. 

35
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% women ($p = 0.03$). This also is consistent with our theory of self-confidence maintenance if women tend to place more weight on anticipatory utility (are more likely to have $\alpha > 0$). Figure 10 plots 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%). This is broadly consistent with the theory, which predicts that subjects will optimally place negative valuations on information when their confidence is below a threshold value $c$. 
Anticipatory utility may then explain the sharp gender differences in valuation curves. Of course, since the relevant $c$ will vary across subjects these relationships are only suggestive.

In sum, there are substantial gender differences in both information processing and information acquisition, and these are consistent with our theory of self-confidence maintenance if women place more weight on anticipatory utility than men.

8 Conclusion

Recent theoretical work has argued that information may affect welfare in more subtle ways than are captured by the traditional paradigm, in which information is useful strictly to improve the accuracy of decision-making. Motivated by this new literature, we build a model to
understand how an agent who prefers to be confident will learn about his own ability. Such an agent reacts less on average to new information than a Bayesian, reacts more to positive than to negative information, and may be averse to obtaining highly informative feedback.

Our experimental design allows us to measure beliefs in an incentive-compatible way and to cleanly separate the role of priors and signals in shaping posterior beliefs. We find that both our predictions regarding updating are borne out in the data: subjects are on average conservative and asymmetric updaters. A substantial fraction also exhibit an aversion to information about their relatively ability which cannot be explained by simple measurement error in reporting. Overall the data support the view that subjects’ careful regulate their self-confidence to the potential detriment of subsequent decision-making.
References


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


Svenson, Ola, “Are we all less risky and more skillful than our fellow drivers?,” Acta Psychologica, February 1981, 47 (2), 143–148.


A Proofs

Proof of Proposition 1:

We define the new random variable $Z^L_T$:

$$Z^L_T = \logit(\mu_T) - \gamma^T_L \sigma^T_L$$

(23)

By the central limit theorem, the distribution of this new random variable – conditional on the agent being a low type – converges in distribution to a normal random variable:

$$Z^L_T \xrightarrow{D} N(0, 1)$$

(24)

Analogously, we can define a random variable $Z^H_T$ that converges to a normal variable conditional on the agent being a high type.

Therefore, the probability that the low type agent’s belief will be above some cost $c$ converges to the probability that a normal random variable is at least $\logit(c) - \gamma^T_L \sigma^T_L$ standard deviations away from its mean which is of order $\exp(-aT)$ for some constant $a > 0$. Similarly, we can bound the probability that the high type makes a mistake.

Proof of Proposition 2:

The agent will invest in the final period iff $\hat{\mu}^T > c$ which implies:

$$\logit \mu + S^T_H \hat{\lambda}_H + S^T_L \hat{\lambda}_L \geq \logit(c)$$

(25)

Hence, for each $(S^T_H, \hat{\lambda}_H, \hat{\lambda}_L)$ there is a critical cost level $c(S^T_H, \hat{\lambda}_H, \hat{\lambda}_L)$ such that the agent is indifferent between investing and not investing. If $T$ is sufficiently large, then $c(T, \lambda_H, \lambda_L) > c(T - 1, \lambda_H, \lambda_L) > \frac{\epsilon}{2}$. A decision-maker who only cares about making the correct investment choice ($\alpha = 0$) will choose $\hat{\lambda}_H$ and $\hat{\lambda}_L$ such that $c(T, \lambda_H, \lambda_L) = c(T, \hat{\lambda}_H, \hat{\lambda}_L)$ and $c(T - 1, \lambda_H, \lambda_L) = c(T - 1, \hat{\lambda}_H, \hat{\lambda}_L)$. Hence, the decision-maker will choose $\hat{\lambda}_H = \lambda_H$ and $\hat{\lambda}_L = \lambda_L$.

Proof of Theorem 1:

We first show conservatism. Assume that the responsiveness of the optimally biased agent, $\max(\hat{\lambda}_H, \hat{\lambda}_L)$, would not converge to zero. In this case, there would be some $\epsilon > 0$ and an infinite sequence $(T_j)$ such that $\max(\hat{\lambda}_H, \hat{\lambda}_L) > \epsilon$. Now consider the mean logit belief $\hat{\gamma}^T_L$ of

22For small $T$ it is possible that less than 2 critical cost levels are contained in the interval $[\epsilon, 1]$. In this case, fewer than 2 equations pin down $\hat{\lambda}_H$ and $\hat{\lambda}_L$ and there are multiple solutions.
the low type, optimally biased Bayesian. There are two possibilities: (a) the sequence \( \hat{\gamma}^{T_j}_L \) converges to negative infinite or (b) the sequence does not converge to negative infinite.

In the first case, the optimally biased Bayesian enjoys lower anticipatory utility than an agent with downward neutral bias – this is a contradiction since downward neutral bias provides a lower bound for the utility of the optimally biased Bayesian. In the second case, there is some lower bound \( M \) and a sub-sequence \((T'_j)\) of \((T_j)\) such that \( \hat{\gamma}^{T'_j}_L > M \). In this case, the low-type agent will take the incorrect action with probability approaching \( \frac{1}{2} \) since the variance of logit-beliefs will be of order \( T \epsilon \). This violates assumption 1.

We next prove asymmetry. Assume that the optimally biased Bayesian is not asymmetric for large \( T \). Then there is a sequence \((T_j)\) such that:

\[
\begin{align*}
\hat{\gamma}^{T_j}_L & \leq \logit(\mu) + T_j \beta^{T_j}_H [q\lambda_H + (1 - q)\lambda_L] \\
\hat{\gamma}^{T_j}_H & = \logit(\mu) + T_j \beta^{T_j}_H [p\lambda_H + (1 - p)\lambda_L]
\end{align*}
\]

There are two possibilities. First of all, \( \hat{\gamma}^{T_j}_H \to \infty \) which implies \( \hat{\gamma}^{T_j}_L \to -\infty \). In this case, the agent enjoys the same anticipatory utility in the limit as the perfect Bayesian. We already know that the downward neutral bias generates higher anticipatory utility which is a contradiction. In the second case, \( \hat{\gamma}^{T_j}_H \to \) does not converge to \(+\infty\). Therefore, there is some \( M \) and some subsequence \((T'_j)\) such that \( \hat{\gamma}^{T'_j}_L < M \). This implies that there is some \( M' \) such that that \( \hat{\gamma}^{T'_j}_L < M' < 0 \). In this case, the downward neutral bias also creates greater anticipatory utility.

**Proof of Proposition 3**

Assume to the contrary that \( \hat{\lambda}_H, \hat{\lambda}_L \) does not converge to \( \frac{1-q}{q} \). In that case, there is some \( \epsilon > 0 \) and a sequence \((T_j)\) such that:

\[
\left| \frac{\hat{\lambda}_H - 1 - q}{\hat{\lambda}_L} \right| > \epsilon
\]

At the same time there has to be an upper bound \( M \) such that:

\[
\logit(\mu) + T \left[ q\hat{\lambda}_H + (1 - q)\hat{\lambda}_L \right] < M
\]

Otherwise, the agent would always make the investment for a sub-sequence of \((T_j)\) which violates the long-run learning condition.

Conditions [27] and [28] together imply that \( \hat{\lambda}_H, \hat{\lambda}_L < \frac{m}{T_j} \) for some constant \( m > 0 \). This implies that (a) the variance of logit-beliefs converge to 0 and (b) there is an upper bound for the mean logit-belief in the high state of the world, \( \hat{\gamma}^{T_j}_H \). However, we can always modify the
basic downward neutral bias introduced in equation (7) and add a small $O(\frac{1}{T})$ drift to $\dot{\lambda}_H$ such that we implement the same $\hat{\gamma}_T^j$ and have $\hat{\gamma}_T^j \to \infty$. This bias would generate higher utility in the high-state of the world because it would reduce the chance of the agent not investing in the good state. This is a contradiction - hence $\hat{\lambda}_H \to \frac{1-q}{q}$.

**Proof of Lemma 1**

The proof is identical to the proof of proposition 1. Since the perfect Bayesian’s error probability converges to zero, she has no need for information in the limit.

**Proof of Theorem 2**

We start with the following lemma.

**Lemma 2** Assume $a > 0$. Then for any $x > 0$ the following holds:

$$\frac{1}{2} (\logit^{-1}(a + x) + \logit^{-1}(a - x)) < \logit^{-1}(a)$$  \hspace{1cm} (29)

To prove this, we calculate $\logit^{-1}(a + x) + \logit^{-1}(a - x) - 2\logit^{-1}(a)$ which has the same sign as:

$$(\exp(x) + \exp(-x))(1 - \exp(a)) + 2\exp(a) - 2$$  \hspace{1cm} (30)

The first term is minimized for $x = 0$. The second term is negative for $a \geq 0$. Hence, the expressions is maximized at $x = 0$ where it is exactly 0. Therefore, it is strictly negative for any $x > 0$. ♣

The above lemma implies that if $\tilde{a}$ is a symmetrically distributed random variable with mean $a$ then $E(\logit^{-1}(\tilde{a})) \leq \logit^{-1}(a)$. Now consider the logit-belief of the low type biased Bayesian. Since $\mu > \frac{1}{2}$ and because logit beliefs at any relative time $\tau$ converge to a normal distribution (which is symmetric), we know that the low type’s anticipatory utility is bounded above by $\alpha(1 - E(c)) \int_T^1 \logit^{-1}(\hat{\gamma}_L^T) d\tau'$. We can approximate this level of anticipatory utility through a downward neutral bias with drift. Hence, the low type’s actual anticipatory utility under the optimal bias converges to this upper bound.

What remains to be shown is that $\hat{\gamma}_L^T \to \zeta$. Assume instead that there is a sequence $(T_j)$ and some $\epsilon > 0$ such that $\logit^{-1}(\hat{\gamma}_L^T) \geq \zeta + \epsilon$. In this case, the low type receives at most extra belief utility of $\alpha(1 - E(c))\epsilon$ but makes mistakes with probability $\frac{1}{1-\zeta}$. Hence the expected loss is at least $\frac{\epsilon}{1-\zeta}$. The agent will prefer $\hat{\gamma}_L^T = \zeta$ provided that:

$$\alpha(1 - E(c))\epsilon \leq \frac{\epsilon}{1-\zeta}$$  \hspace{1cm} (31)

This implies $\alpha \leq \frac{\epsilon}{(1-\zeta)(1-E(c))}$ which is implied by the long-run learning condition.
Proof of Proposition 4
Under the assumption of theorem 2 the naive Bayesian willingness to pay satisfies:
\[
\lim_{T \to \infty} WTP(x, \tau | 0, \mu, \lambda, \lambda) = WTP^S(x)
\] (32)

B Are Negative Valuations Attributable to Noise?
If subjects are not careful recording their answers there may be cases where they record a lower value for $2 and information than for $2 alone simply by chance. This note constructs a formal test of this null hypothesis under (weak) assumptions about the structure of reporting errors.

Let \( S_i, S_i + C_i, \) and \( S_i + P_i \) be agent \( i \)'s true valuation of $2, $2 and coarse feedback, and $2 and precise feedback, respectively. Drop \( i \) subscripts for brevity. We assume that agents report these quantities with additive errors that are distributed normally, identically, independent of each other, and independent of true valuations, so that we observe

\[
\begin{align*}
\hat{S} &= S + \epsilon_S \\
\hat{C} &= S + C + \epsilon_C \\
\hat{P} &= S + P + \epsilon_P
\end{align*}
\]

where \( \epsilon_z \sim N(0, \sigma_z^2) \) for \( z \in \{S, C, P\} \). The second moments of our data are

\[
\begin{align*}
V(\hat{S}) &= V(S) + \sigma^2 \\
V(\hat{C}) &= V(S) + V(C) + 2Cov(S, C) + \sigma^2 \\
V(\hat{P}) &= V(S) + V(P) + 2Cov(S, P) + \sigma^2 \\
Cov(\hat{S}, \hat{C}) &= V(S) + Cov(S, C) \\
Cov(\hat{S}, \hat{P}) &= V(S) + Cov(S, P) \\
Cov(\hat{C}, \hat{P}) &= V(S) + Cov(S, C) + Cov(S, P) + Cov(C, P)
\end{align*}
\]

This system is not point-identified as there are 7 parameters and 6 equations. However, we can bound the parameters by imposing the requirements that variances be positive and correlation coefficients within \([-1, 1]\). To bound \( \sigma^2 \) note that

\[
\begin{align*}
V(C) &= V(\hat{C}) + V(\hat{S}) - 2Cov(\hat{S}, \hat{C}) - 2\sigma^2 \\
V(P) &= V(\hat{P}) + V(\hat{S}) - 2Cov(\hat{S}, \hat{P}) - 2\sigma^2 \\
Cov(C, P) &= Cov(\hat{C}, \hat{P}) + V(\hat{S}) - Cov(\hat{S}, \hat{C}) - Cov(\hat{S}, \hat{P}) - \sigma^2
\end{align*}
\]
which implies the following must hold:

\[
\sigma^2 \leq \frac{1}{2} \left( V(\hat{C}) + V(\hat{S}) - 2\text{Cov}(\hat{S}, \hat{C}) \right)
\]

\[
\sigma^2 \leq \frac{1}{2} \left( V(\hat{P}) + V(\hat{S}) - 2\text{Cov}(\hat{S}, \hat{P}) \right)
\]

\[
-1 \leq \frac{\left( \text{Cov}(\hat{C}, \hat{P}) + V(\hat{S}) - \text{Cov}(\hat{S}, \hat{C}) - \text{Cov}(\hat{S}, \hat{P}) - \sigma^2 \right)}{\sqrt{V(\hat{C}) + V(\hat{S}) - 2\text{Cov}(\hat{S}, \hat{C}) - 2\sigma^2}} \leq 1
\]

The largest value of \( \sigma \) that satisfies these restrictions for our data is \( \sigma \approx 26.4 \). Now fix any

Figure 11: Noise Tests

Plots probabilities of observing \( n(x) \) reported information values less than \( x \) under the null hypothesis that all true information values are 0, for various values of \( x \).

\( x < 0 \) and let \( n(x) \) be the number of observations for which both \( \hat{C}_i - \hat{S}_i < x \) and \( \hat{P}_i - \hat{S}_i < x \). Under the null hypothesis that \( C_i \) and \( P_i \) are bounded below by 0, the probability that these inequalities hold for any agent \( i \) is at most \( \zeta(x, \sigma^2) \equiv P(\epsilon_S \geq \max\{\epsilon_C, \epsilon_P\} - x) \), the probability when \( C_i = P_i = 0 \). This can be calculated numerically for any given \( x \) and \( \sigma^2 \), and consequently the probability that \( \hat{C}_i - \hat{S}_i < x \) and \( \hat{P}_i - \hat{S}_i < x \) hold for \( n(x) \) or more out of \( N \) individuals
in a sample can be bounded by

\[
p(x, \sigma^2) \equiv \sum_{m=n(x)+1}^{N} \binom{N}{m} \zeta(x, \sigma^2)^m (1 - \zeta(x, \sigma^2))^{N-m}
\]  

(33)

We calculated \( p(x, \sigma^2) \) for \( \sigma = 26.4 \) and for a variety of thresholds \( x \). Figure plots the results. For any threshold below \(-60\) we can reject the null at the 0.01 level.

C Additional Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>By Quiz Type</th>
<th>Overall</th>
<th>By Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
<td>Score</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( N )</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>Restricted Sample</td>
<td>656</td>
<td>10.2</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Full Sample</td>
<td>1058</td>
<td>9.7</td>
<td>4.3</td>
</tr>
<tr>
<td>By Quiz Type</td>
<td>Correct</td>
<td>Incorrect</td>
<td>Score</td>
<td>Correct</td>
</tr>
<tr>
<td>1</td>
<td>79</td>
<td>8.1</td>
<td>3.1</td>
<td>1.7</td>
</tr>
<tr>
<td>2</td>
<td>85</td>
<td>13.0</td>
<td>2.9</td>
<td>2.7</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>8.9</td>
<td>3.3</td>
<td>3.0</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>12.2</td>
<td>3.8</td>
<td>3.1</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>6.5</td>
<td>1.6</td>
<td>4.0</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>14.5</td>
<td>4.5</td>
<td>2.3</td>
</tr>
<tr>
<td>7</td>
<td>73</td>
<td>7.6</td>
<td>2.6</td>
<td>2.2</td>
</tr>
<tr>
<td>8</td>
<td>69</td>
<td>13.6</td>
<td>2.8</td>
<td>3.2</td>
</tr>
<tr>
<td>9</td>
<td>69</td>
<td>7.3</td>
<td>3.5</td>
<td>2.7</td>
</tr>
<tr>
<td>By Gender</td>
<td>Correct</td>
<td>Incorrect</td>
<td>Score</td>
<td>Correct</td>
</tr>
<tr>
<td>Male</td>
<td>314</td>
<td>10.6</td>
<td>4.2</td>
<td>2.7</td>
</tr>
<tr>
<td>Female</td>
<td>342</td>
<td>9.7</td>
<td>4.4</td>
<td>2.8</td>
</tr>
</tbody>
</table>