No one to blame: Biased belief updating without attribution*†

Alexander Coutts Leonie Gerhards Zahra Murad

February 19 2019

Abstract

Individuals are often overconfident about their ability, affecting career and financial decisions. We investigate how overconfidence persists in the face of objective feedback. Self-attribution biases exist when we take credit for good outcomes, but blame poor outcomes on external factors. While heavily studied in social psychology, and often referenced in economics, rigorous evidence is scarce. We present a modified Bayesian model of self-attribution bias, which distinguishes biases in attribution towards idiosyncratic noise versus a stable fundamental factor. In an experiment where individuals receive noisy performance feedback that also depends on a teammate, we identify precise patterns in attribution among these two dimensions of uncertainty. Individuals are overconfident and update in a biased self-serving way relative to the Bayesian benchmark and a control group. Fundamental attribution bias is rejected, as negative feedback is disproportionately attributed to noise rather than the teammate. Yet self-serving biases also spillover to inflate beliefs about the teammate. This suggests important implications of the consequences of biased information processing, beyond the direct effect of overconfident beliefs about oneself.

*Coutts: Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, Rua da Holanda 1, 2775-405 Carcavelos, Portugal (email: alexander.coutts@novasbe.pt); Gerhards: Department of Economics, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany (e-mail: leonie.gerhards@wiso.uni-hamburg.de); Murad: Economics and Finance, The University of Portsmouth, Portsmouth, PO1 2UP, United Kingdom (email: zahra.murad@port.ac.uk).

†We are very grateful for useful comments from seminar and conference participants at University of Alicante, University of Amsterdam, ESA Berlin, HEC Lausanne, Lisbon Game Theory Meetings, M-BEES, NYU CESS, NYU Shanghai, University of Portsmouth, SHUFE, THEEM, and WZB.

1 Introduction

Overconfidence has been shown to be a persistent bias in human decision making, and has been linked to financial decision making (Barber and Odean, 2001), CEO investment decisions (Malmendier and Tate, 2005), as well as career choice (Köszegi and Rabin, 2006). The persistence of overconfidence is especially puzzling when considering that individuals receive informative but imperfect feedback about their ability in many contexts. In this paper we focus on how overconfidence may persist through biased information processing about self-relevant information, particularly when this information may come bundled with an additional source of uncertainty.[1] For example, consider a student who receives a grade for group work, an employee who receives a bonus based on her team's performance, or a trader who realizes a return based on her portfolio and the underlying state of the economy.

Why might individuals process self-relevant information in a biased manner? A large literature in psychology is dedicated to the study of self-attribution bias, which involves an over-attribution of past successes to internal factors such as ability, relative to failures which are attributed to external factors. For example, the student above would be biased if he takes credit for high grades, but blames his colleagues for low grades. The underlying motivation for such behavior is often traced back to Freudian principles: the pleasures associated with success and the pains associated with failure (Weiner and Graham, 1999). Such motivated cognition thus can enhance pleasure and/or reduce pain through biased attribution patterns.[2]

A distinguishing feature of these examples, and indeed of studies of self-attribution bias in social psychology beginning with Heider (1958), are that attribution typically occurs vis-a-vis some external fundamental: e.g. the state of the economy or one's colleagues, as in our previous examples. Alternatively, one could attribute outcomes to idiosyncratic factors such as chance, e.g. the student above blaming bad luck for the group's low grade. Different forms of self-attribution bias can have vastly different consequences. The way we process information about ourselves can affect our views on other people or

[1] Importantly, we do not examine other channels involving assessment of past or future information, such as biased memory or selective information acquisition. Considering the past, hindsight-bias or biased memory theories, see Fischhoff (1975) and Benabou and Tirole (2002), could lead to overconfidence if individuals recall information in a biased way; Zimmermann (2017) in fact find evidence of asymmetric recall of feedback. Regarding the future, individuals may selectively sample information, choosing only sources of information that are likely to nurture overconfidence, e.g. Eliaz and Spiegler (2006).

[2] There could also be self-motivational or signalling motives for ego-enhancement or protection, see Benabou and Tirole (2002). As Weiner and Graham (1999) note, there may also be cognitive explanations, though the overall evidence casts some doubt on these. Our experimental results also rule out such cognitive explanations.

factors, and thus shape our future decisions. A student with a failing group grade will try and find a new study group if he blames his colleagues, but will give them another chance if he blames luck.

In our theory and corresponding experiment, we distinguish between idiosyncratic (unstable) factors, and dispositional (stable) factors. Most economic research on belief updating about own ability has studied only attribution towards idiosyncratic factors: individuals form a prior, receive noisy signals, and then it is observed how they attribute feedback between their own ability and the noisy signal, e.g. Möbius et al. (2014).

In our paper there is a critical role played by a second dimension of uncertainty: stable but unknown factors, such as the underlying fundamental of the state of the economy or the quality of colleagues. We present a modified model of Bayesian updating which leverages these two dimensions of uncertainty to test two types of self-attribution bias. In doing so we operationalize a concept that has been the subject of hundreds of psychological studies, but has until now lacked a quantitative, falsifiable, model.[3] Unlike existing work, our focus is on specifying a simple but tractable model of self-attribution bias, and directly testing it. We thus provide discipline on existing theoretical and empirical work which appeals to self-serving attribution bias.[4]

The first type, noisy attribution bias (NAB), prescribes that individuals over- (under-) attribute negative (positive) feedback to noise, but otherwise accurately process information about states of the world (external fundamentals) not directly relevant to their self-image. This can be interpreted as a model where individuals systematically misinterpret the precision of the signal depending on whether news is good or bad. This type of bias has been studied in previous studies of belief updating (Möbius et al., 2014; Buser et al., 2018; Coutts, 2018; Eil and Rao, 2011).

The second, fundamental attribution bias (FAB), posits that individuals respect the aggregate precision of the signal, but make errors in its attribution. In particular, they tend to over-attribute positive outcomes to themselves, but negative outcomes to an external fundamental. This type of bias has not been carefully studied empirically, as it necessitates observing belief updates for two-dimensions of uncertainty.

Determining the extent of attribution biases is crucial for our understanding of decision making under risk, yet distinguishing NAB from FAB has not been possible in

[3]This statement is not intended to disparage research on attribution-biases in social psychology, as these studies have contributed greatly to our understanding of these biases, and form the motivation for this paper. Rather it reflects differences in methodological approaches, namely the social psychological theories regarding attribution biases are typically non-quantitative theories.

[4]Existing work includes Billett and Qian (2008), Daniel et al. (1998a), Doukas and Petmezas (2007), Gervais and Odean (2001), Hilary and Menzly (2006), Hoffmann and Post (2014), Kim (2013), Li (2010), and Libby and Rennekamp (2011).

previous work, despite important implications. Self-serving biases not only have direct effects on beliefs about self, but also may affect our view of others which can further impact future decisions. As a first benchmark, our theoretical model of both NAB and FAB predicts asymmetric updating in personally relevant contexts. But FAB imposes a further penalty since the decision maker will forecast both own ability and the external fundamental with error. A student may choose new partners, or a trader might try new markets, against their own self-interest if she suffers from FAB. An extreme example of such self-defeating learning is a focus of Heidhues et al. (2018), where a decision maker is caught in a vicious circle of taking ever worsening decisions and having increasingly pessimistic assessments about an external fundamental.

We test these two theories of self-attribution bias using a natural context whereby a two-person team's output depends on the ability of both members. Individuals receive aggregate team feedback and must attribute the feedback to both their own and their teammate's ability. The updating problem is then one of joint inference - an individual must update beliefs about his own ability, and the ability of the teammate. The motivation for a realistic context involving another person comes directly from the psychology of attribution biases. Early work in this literature posited that people are naturally inclined to make attributions to more stable or dispositional factors, which includes traits of individuals, rather than unstable or situational factors, Heider (1958).[5] Similar arguments can be made appealing to availability bias, Tversky and Kahneman (1973), which would also suggest that attribution will be towards factors which come easier to mind.

In the experiment we study the evolution of beliefs about ability: here scoring in the top 50% of a reference group on an IQ-style test. We elicit prior beliefs about both teammates' performances, as well as posterior beliefs each time feedback is received. We use a novel, passive elicitation procedure that is incentive compatible, invariant to risk attitudes, but also intuitive. To implement this procedure we utilize a concave mapping between a chosen weight on own vs. teammate performance, and a probability of earning a payoff of €10. To maximize the probability of earning the €10, subjects need to form a subjective assessments of their and their teammate's ability, and calculate the optimal weight. Our elicitation procedure provides them the opportunity to avoid this complex optimization problem, by simply entering their subjective beliefs, and based on these, having the computer show them the true probability of earning €10 for every possible weight combination in $[0, 1]$. In this way, whenever they prefer a higher probability of earning €10,

[5]The psychology of this relates also to the basic human desire to seek meaning in and understand the world, Shaver (1983). An example is Spilka et al. (1985) who note that the existence of God and religion can be interpreted as a manifestation of such a desire for meaning.

they should truthfully enter their beliefs to maximize their expected earnings.

One final but critical component is that we conduct the entire experiment with a high-powered control group. Given previous work which casts doubt on using Bayes' rule as a benchmark, we had approximately half of our subjects participate in the same experiment, but acting as a third party and taking decisions for a random stranger, in the same team format, about the stranger's test performance rather than their own. That is, the experiment is identical, but the subject's own ability, and hence ego, is no longer relevant to the decision problem. In this way we are able to compare belief formation and updating to both the Bayesian benchmark, as well as the benchmark generated by our non-ego relevant control.

Our main outcomes of study are: (i) the formation of prior beliefs, (ii) the mechanics of the updating process, and (iii) final posterior beliefs. To study (ii) we use a structural framework which nests Bayes' rule as a special case, while for (iii) we use a non-parametric strategy to match subjects reporting beliefs about own performance with subjects with the same priors about others in our control. Our results are as follows. First, we find strong evidence that individuals are overconfident, following tests based on Dubra and Benoît (2011). Next, we find that individuals attribute team feedback in different ways to themselves and their teammate. In particular, they over-attribute positive feedback to themselves, while significantly under-weighting negative feedback for both teammates. While this latter result runs contrary to both types of self-attribution biases, it suggests a form of attribution to noise which results in biased updating about both team members. In our control, we do not observe any evidence of biased attribution of positive or negative feedback. Finally, using our matching strategy we show that final posterior beliefs about own performance are significantly higher after feedback compared with beliefs about another's performance, for the same initial prior beliefs. In line with the attribution findings, there is evidence that these differences are greater for subjects who received more negative signals, with subjects receiving all negative signals holding significantly greater beliefs about both teammates, relative to the control.

Overall our results suggest that overconfidence biases may be nurtured through biased processing of information. While we reject the mechanics of the two forms of attribution bias specified, the results tell a clear story of biased attribution which results in inflated beliefs about one's self and one's teammate. This suggests that overconfidence is persistent, but also that self-serving information processing can spillover to bias estimates of other sources of uncertainty. This has key implications beyond simple mis-characterization of noise, since it suggests that individuals will form biased expectations about other states of the world as a consequence of their desire to preserve their own positive self-image.

2 Related Literature

Our study links multiple strands of literature in economics and psychology, namely: those on overconfidence, attribution biases, and belief updating. Behavior consistent with overconfidence about ability has been documented in numerous settings, such as driving (Svenson, 1981), financial trading (Barber and Odean, 2001), as well as in a number of lab experiments concerning tests of academic ability. While Dubra and Benoît (2011) noted that rational data may generate overconfident-appearing data, even accounting for this many studies have found evidence consistent with overconfidence, see Benoît et al. (2015), and discussion contained therein.[6]

How have overconfident beliefs documented in these studies persisted in the face of opportunities for learning and receiving feedback about one's ability? An important explanation comes from the theory of self-serving or self-attribution bias: the tendency to "attribute success to our own dispositions and failure to external forces" (Hastorf et al., 1970). Individuals with self-attribution bias tend to ascribe their successes to internal factors such as their skill, ability, effort, and personality, while ascribing their failures to external factors such as luck, other's actions, and institutional environments.

Self-attribution bias has its origins in the writings of Fritz Heider. Heider (1958) described how people have an innate desire to explain behaviors and outcomes. He described the concept of causal attribution, the "naive psychology" through which individuals try to understand behaviors and outcomes, as a result of some underlying causal process. Heider (1944) noted that people tend to attribute outcomes to more salient sources such as other individuals, rather than objects or luck, with clear parallels to availability bias of Tversky and Kahneman (1973).[7]

The resulting theories of attribution are focused more on general principles rather than tractable models, as discussed in Kelley (1973) and Weiner (2010). Most of the literature has been empirical. In an archetypal experiment in social psychology studying self-attribution in achievement tasks, individuals are given a task (e.g. an anagrams task), then receive success or failure feedback (sometimes falsified), and next asked to distribute responsibility for this outcome among internal vs external factors (Miller and Ross, 1975; Mezulis et al., 2004). While these studies are provoking, it is also difficult to understand

[6]The focus of these studies is on over-placement relative to others, see Moore and Healy (2008) for distinctions between types of overconfidence. Of note is that some studies have not found data which appear overconfident in relative comparisons. For example, prior beliefs in Möbius et al. (2014) do not show evidence for overconfidence. However, to our knowledge, no high-powered studies have found aggregate underconfidence in relative performance assessments.

[7]Empirical evidence on this has can be found in Pryor and Kriss (1977) with further discussion in Lassiter et al. (2002).

how subjects will interpret some of these concepts, and how they should respond to feedback that may have been falsified.[8] Early meta-analyses were conducted by Miller and Ross (1975), Zuckerman (1979), and Arkin et al. (1980). Miller and Ross (1975) found only evidence of attribution biases for success but not for failure, and attributed this more towards cognitive bias. However later studies found evidence in both success and failure, as reported in a large meta-analysis by Mezulis et al. (2004).

Bridging this theory and evidence within social psychology with the prevailing quantitative theories in economics presents some challenges. The identity of the source of uncertainty is crucial in psychology, but in fact is irrelevant in a Bayesian framework. Our focus is on making sure core principles from psychology do not get lost in their translation to a Bayesian framework. Our contribution in this paper is to quantify the self-attribution bias framework, and we intentionally focus our theoretical discussion and experiment on having a human teammate's ability to serve as a salient source of uncertainty. In addition, in our experiment subjects remain matched and receive team feedback for multiple rounds, creating a source of uncertainty that is likely to be perceived as more stable, which is directly in line with the motivations of self-attribution bias.[9]

Our focus on belief updating with two dimensions of uncertainty puts us in the middle of two literatures in economics: a recently growing literature on motivated cognition and belief updating involving one dimension of uncertainty, and an emerging literature on learning and decision making with multiple dimensions of uncertainty.[10] This former strand of literature includes work by Möbius et al. (2014), Buser et al. (2018), Coutts (2018), Grossman and Owens (2012), Ertac (2011), and Eil and Rao (2011). These authors focus primarily on capturing reduced form aspects of asymmetric information processing about personal qualities, which may also be consistent with the predictions of self-attribution bias.[11] Möbius et al. (2014) present a theory which provides a common motivation for this line of research, a model of asymmetric updating bias that arises from a world where individuals derive direct utility from believing they have high ability, à la

[8] See Pekrun and Marsh (2018) for a more detailed discussion of some empirical concerns of this literature. As Silvia and Duval (2001) note, some concepts such as "luck" are not straightforward to interpret outside of a quantitative framework. Additionally, it is unclear whether blaming external factors for falsified failure, if one is convinced they did not fail, can be interpreted as bias.

[9] Re-matching individuals with new teammates every period would reduce stability, and would be predicted to result in less attribution.

[10] In addition, our experiment can be seen as a validation of recent applications of self-attribution to financial markets or trader behavior, see Daniel et al. (1998b) and Gervais and Odean (2001), as well as our references in the introduction. Regarding these empirical studies, it is difficult to establish causality as investors or managers are not randomly assigned outcomes. Our paper contributes to identifying the scope for these biases in a controlled environment.

[11] Other authors have studied updating about non-personal qualities such as financial stakes (Barron, 2017) as well as about qualities of others (Erkal et al., 2019).

Brunnermeier and Parker (2005).[12] However this literature is not well equipped to study the mechanics of self-attribution biases, since attribution can only be to one source, noise, by construction.

This is important because most real world updating problems involve more than one source of uncertainty. Turning now to the latter strand of literature within economics, we discuss two relevant theoretical studies, Heidhues et al. (2018) and Hestermann and Le Yaouanq (2018).[13] Both study the long run consequences of confidence biases for decision making with two dimensions of uncertainty, ability and another external fundamental, assuming Bayesian updating. In contrast our focus is on short term updating biases. However, it is worth discussing some overlap with each paper in turn, when possible we discuss their theory in our context of teams.

Hestermann and Le Yaouanq (2018) study the consequences of initial mis-calibration in confidence in a world where individuals are matched with some fundamental but can change their environment, i.e. match with a new fundamental at some cost. Initially overconfident individuals rationally attribute successes as reflective of their ability, while they attribute failures as reflective of the fundamental.[14] Our experimental setup relates to their theory, as our feedback structure is a particular case of their setup, where there is neither complementarity nor substitutability between teammates' abilities.

Unlike Hestermann and Le Yaouanq (2018), Heidhues et al. (2018) assume that individuals believe with certainty that their ability is higher than it really is, and remain matched to a constant underlying fundamental.[15] They demonstrate that under certain conditions, an overconfident individual will perceive poor outcomes as reflecting poor performance by other team members rather than herself. In response, they show that

[12]Evidence of asymmetric information processing is mixed, see Benjamin (2019). Positive asymmetry (Eil and Rao, 2011; Möbius et al., 2014), no asymmetry (Grossman and Owens, 2012; Buser et al., 2018), and negative asymmetry (Coutts, 2018; Ertac, 2011) have all been observed. Buser et al. (2018) do find positive asymmetry in some sub-samples. Reactions to feedback have also been studied in less comparable settings, see Burks et al. (2013), Eberlein et al. (2011), Pulford and Colman (1997), Ertac and Szentes (2011), and Wozniak et al. (2016).

[13]A related theoretical paper is Deimen and Wirtz (2016), who examine the optimal strategy of an agent who faces two dimensional uncertainty: own ability, and the returns to effort in the environment she faces. They find heterogeneity in the optimal strategy depending on the costs of effort as well as on initial beliefs about ability.

[14]There are asymmetric dynamic consequences of initial biases in confidence: overconfident individuals end up being dissatisfied with their environment (and hence quit "too early"), while initially underconfident individuals are more likely to be satisfied with the environments they find themselves in, and hence may remain "stuck".

[15]That is, the environment is unchangeable. Regarding the overconfidence assumption, they take steps to show how it can be relaxed, by considering a form of biased updating and showing that this does not change the core predictions of their theory. In this extended framework individuals receive continuous signals about ability which are biased upwards by a fixed amount. This differs in both scope and consequence from our theory.

the individual decision making can lead to a cycle of self-defeating learning, and poor outcomes which the agent increasingly attributes to her other teammates.

Our setup is a variation of both these models, though with the crucial difference that we study non-Bayesian information processing due to self-attribution bias. Like Heidhues et al. (2018), our decision involves a delegation-type decision between two teammates. However, in our environment we shut-down the feedback mechanism from this decision, which precludes the type of self-defeating learning they study. In our setup, these dynamics can only occur through the channel of biased inference, not through the link between delegation decisions and outcomes.

3 Theory

3.1 Preliminaries

Our theory describes a situation where a decision maker is faced with two sources of uncertainty, and must update beliefs in order to take a decision. We refer to these sources explicitly as the ability of teammate 1 and the ability of teammate 2, which facilitates comparison with the experiment.

We consider finite time periods $t \in \{1, 2, ..., \tau\}$. We further restrict our attention to a discrete 2×2 state space, again which will correspond to the experiment. Teammate 1's unknown ability is given by $a \in \{B, T\}$, corresponding to either low ability (bottom half of performance distribution) or high ability (top half). The unknown fundamental of interest $\phi \in \{B, T\}$ is defined analogously, where in the experiment this will correspond to whether teammate 2 is in the bottom half or top half of performances respectively.

The state space therefore consists of four states, it will be convenient to denote these in short hand:

$$S_1 S_2 = \begin{cases} TT & \text{if } a = T \text{ and } \phi = T \\ TB & \text{if } a = T \text{ and } \phi = B \\ BT & \text{if } a = B \text{ and } \phi = T \\ BB & \text{if } a = B \text{ and } \phi = B \end{cases}$$

At each time period, decision makers take an action, by choosing how much to weight the performance of teammate 1 relative to teammate 2, ω_t. It will become clear what the incentives are behind taking this weighting decision after we describe the structure of payoffs. Monetary payoffs at time t, $\Pi^t(\omega_t, a, \phi)$, are awarded probabilistically, with

the possibility of earning a payment $P > 0$ or nothing. The individual will optimize by considering the payoffs of each period, which are determined according to the following lottery. $(P, 0; \sqrt{\omega_t})$ is the lottery that pays P with probability $\sqrt{\omega_t}$ and 0 otherwise.

$$\Pi^t(\omega_t, a, \phi) = \begin{cases} P & \text{if } TT \\ (P, 0; \sqrt{\omega_t}) & \text{if } TB \\ (P, 0; \sqrt{1-\omega_t}) & \text{if } BT \\ 0 & \text{if } BB \end{cases} \tag{1}$$

3.2 Optimal weight

We assume that individuals are subjective expected utility maximizers, with utility function $u(\cdot)$ which is continuous, strictly increasing, and differentiable. Individuals form subjective beliefs about the probabilities that teammate 1 and 2 are in the top half. Thus, agents have priors $\tilde{a}_0$ and $\tilde{\phi}_0$ about the probability that $a = T$ and $\phi = T$ at time $t = 0$ respectively.

Denote beliefs about the four states at time t by: $b_t^{S_1 S_2}$. For brevity, it will often be convenient to represent beliefs about all four states as a 4×1 vector, $\mathbf{b}_t$. At time $t = 0$, before receiving any feedback, the prior beliefs $\tilde{a}_0$ and $\tilde{\phi}_0$ are independent. Thus $b_0^{TT} = \tilde{a}_0 \cdot \tilde{\phi}_0$, and so on. The optimization problem of individuals is to maximize expected utility, $Q(\omega_t, \mathbf{b}_t)$:

$$\begin{aligned} Q(\omega_t, \mathbf{b}_t) = b_t^{TT} \cdot u(P) \\ + b_t^{TB} \cdot \sqrt{\omega_t} \cdot u(P) + b_t^{TB} \cdot (1 - \sqrt{\omega_t}) \cdot u(0) \\ + b_t^{BT} \cdot \sqrt{1-\omega_t} \cdot u(P) + b_t^{BT} \cdot (1 - \sqrt{1-\omega_t}) \cdot u(0) \\ + b_t^{BB} \cdot u(0) \end{aligned} \tag{2}$$

Taking first order conditions and setting the resulting equation equal to 0:

$$b_t^{TB} \cdot \frac{1}{2\sqrt{\omega_t}} \cdot [u(P) - u(0)] = b_t^{BT} \cdot \frac{1}{2\sqrt{1-\omega_t}} \cdot [u(P) - u(0)] \tag{3}$$

This leads to the optimal weight,

$$\omega_t^* = \frac{1}{1 + \left(\frac{b_t^{BT}}{b_t^{TB}}\right)^2}. \tag{4}$$

Note that the optimal weight does not depend on the curvature of the utility function, and hence is independent of risk preferences. Note further, that intuitively, the optimal weight ω_t^* is increasing in b_t^{TB}, the belief that teammate 1 is in the top half and teammate 2 is in the bottom half, and is decreasing in b_t^{BT}, the belief that teammate 2 is in the top half and teammate 1 is in the bottom half. Further, unless there is certainty, extreme weights are never optimal.

We now pause to note a few things. First, given the functional form of our payoff function, $Q(\cdot)$, we note that the optimum in Equation 4 is guaranteed to exist, and is unique for any beliefs except for the extreme case when $b_t^{TB} = b_t^{BT} = 0$.[16] Next, in period 0, this functional form generates precisely the sufficient condition which would guarantee self-defeating learning in Heidhues et al. (2018). The optimal weight depends in opposite directions on the expected ability of the individual and the expected ability of teammate 2.[17] In our setup, the feedback that our agents receive is independent of their weighting decisions. Thus while overconfidence and potentially biased updating in our context reduces expected payoffs, learning cannot be self-defeating in the sense of Heidhues et al. (2018).

3.3 Feedback

We have described payoffs, and determined the optimal weight ω_t, given beliefs $\mathbf{b}_t$. We now discuss binary feedback, which is received after every period, and importantly, is independent of chosen weights ω_t in previous periods, and independent of past feedback.

Binary signals $\mathcal{S}$ are independent across time and we denote them in shorthand by $\mathcal{S} = (1, 0; \Phi_{S_1 S_2})$. Signals take the value 1 with probability $\Phi_{S_1 S_2}$ and 0 with probability

[16] Note that when $b_t^{TB} = 0$ and $b_t^{BT} > 0$, the unique optimal weight is $\omega_t^* = 0$. In the extreme case where both $b_t^{TB} = 0$ and $b_t^{BT} = 0$, payoffs are identical for every possible weight. Hence any weight is optimal. By the laws of probability $b_t^{TB} + b_t^{BT} \leq 1$.

[17] Heidhues et al. (2018) have a continuous state space for ability, while ours is binary. Thus, to be certain about ability and overconfident in our setting reduces to $a = 1$. To see the result on self-defeating learning, note that one can rewrite $Q(\omega_t, \mathbf{b}_t)$ in terms of priors about the ability of teammate 1 $\tilde{a}$ and teammate 2 $\tilde{\phi}$. Then one can see that: $Q_{\tilde{a}}(\cdot) > 0$, $Q_{\tilde{\phi}}(\cdot) > 0$, $Q_{\omega\tilde{\phi}}(\cdot) < 0$, and $Q_{\omega\tilde{a}}(\cdot) > 0$. In other words, expected utility is increasing in expected ability of teammate 1 and 2, $\tilde{a}$ and $\tilde{\phi}$ respectively, and the optimal weight ω* is decreasing in the expected ability of teammate 2 $\tilde{\phi}$ and increasing in expected ability of teammate 1 $\tilde{a}$.

$1 - \Phi_{S_1 S_2}$. Given this feedback structure, we now describe a theory of biased information processing which follows the mechanics of Bayes rule but generates biases in attribution of the form discussed in the psychology and economics literature.

3.4 Belief Updating

Using Bayes' rule one can calculate how beliefs evolve for the four states, and hence how beliefs about being in the top half evolve. Note that $b_t^1 = b_t^{TT} + b_t^{TB}$, that is the belief that teammate 1 is in the top half is equal to the sum of the beliefs about the probability of states TT and TB. A Bayesian will update beliefs about teammate 1 being in the top half given either a signal of 1 or 0 respectively as follows:[18]

$$\begin{aligned}
\left[b_{t+1}^{1,BAYES} | \mathcal{S} = 1\right] &= \frac{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}}{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} T b_t^{BB}} \\
\left[b_{t+1}^{1,BAYES} | \mathcal{S} = 0\right] &= \frac{(1 - \Phi_{TT}) b_t^{TT} + (1 - \Phi_{TB}) b_t^{TB}}{(1 - \Phi_{TT)} b_t^{TT} + (1 - \Phi_{TB}) b_t^{TB} + (1 - \Phi_{BT}) b_t^{BT} + (1 - \Phi_{BB}) b_t^{BB}}.
\end{aligned} \tag{5}$$

Analogously for teammate 2, where $b_t^2 = b_t^{TT} + b_t^{BT}$:

$$\begin{aligned}
\left[b_{t+1}^{2,BAYES} | \mathcal{S} = 1\right] &= \frac{\Phi_{TT} b_t^{TT} + \Phi_{BT} b_t^{BT}}{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} T b_t^{BB}} \\
\left[b_{t+1}^{2,BAYES} | \mathcal{S} = 0\right] &= \frac{(1 - \Phi_{TT}) b_t^{TT} + (1 - \Phi_{BT}) b_t^{BT}}{(1 - \Phi_{TT)} b_t^{TT} + (1 - \Phi_{TB}) b_t^{TB} + (1 - \Phi_{BT}) b_t^{BT} + (1 - \Phi_{BB}) b_t^{BB}}.
\end{aligned} \tag{6}$$

3.5 Self-Attribution Bias

Recalling the literature on attribution biases more generally, psychologists such as Heider drew important distinctions between attributions to other people versus objects, and from stable qualities such as the ability of oneself or another, versus unstable qualities such as randomness or luck. In describing self-attribution bias, the focus is on mis-attribution to self relative to other factors.

In this section we quantify these ideas into a theory of Bayesian updating which maintains the structure of Bayes' rule but allows for mis-attribution of feedback. In our context there are three possible sources to attribute feedback to: (1) performance of teammate 1; (2) performance teammate 2; (3) noise. Noise is present since signals are in general not

[18] To derive this equation note that the probability of $\mathcal{S} = 1$ conditional on being in the top half is $\frac{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}}{b_t^1}$. The probability of being in the top half is, b_t^1, and the probability of receiving any signal $\mathcal{S} = 1$ is $\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}$.

perfectly informative about the states of the world, i.e. $\Phi_{S_1,S_2} \in (0,1)$. Following previous work, we refer to the teammates' performance as the two dimensions of uncertainty.

For biased attribution to arise there must be some benefit for holding positive beliefs about one of these dimensions. We highlight the case where the decision maker herself is teammate 1, as in part of the experiment. According to the literature in psychology, we would expect an asymmetric self-attribution bias when an agent herself is teammate 1. Thus, the key focus of mis-attribution will relate to inflating beliefs about factor (1) at the expense of factors (2) and (3). There are a number of potential explanations for this. Individuals could desire to enhance their self-esteem or protect their ego, consistent with a model where they receive instrumental utility from positive beliefs (Möbius et al. (2014)); overconfidence may benefit internal motivation (Benabou and Tirole (2006)); or it may serve social signalling purposes (Burks et al. (2013)).[19] Importantly, we are agnostic over the underlying source of this utility.

Given these two dimensions of uncertainty, we are uniquely able to present a test of two models of non-Bayesian motivated attribution errors which bias beliefs about ones self in the positive direction. A Bayesian attributes proportionately among these three sources. In the first model, noisy attribution bias (NAB), the agent processes accurate information about other factors (teammate 2) but is positively biased about own performance (teammate 1) at the expense of noise. Thus in NAB, the agent mis-attributes between (1) and (3), but attributes (2) correctly. In the second model, fundamental attribution bias (FAB), the agent respects the amount of noise contained in the signal, but is biased about her own performance (teammate 1) at the expense of teammate 2. Thus she mis-attributes between (1) and (2), but attributes (3) correctly. While it is interesting to consider a more flexible model of mis-attribution across these three sources, for the purposes of generating clear empirical benchmarks, our focus was on these more stark conjectures. Our experiment thus presents a clear way to distinguish NAB from FAB.

The psychology literature suggests that we should expect that the target of this mis-attribution is more likely to be the other teammate (2) rather than noise (3), which is a unique feature of FAB. It is worth noting that the extent and form of self-attribution biases may themselves be a function of the environment. That is the activation of FAB or NAB could be context specific, depending on whether there is or is not an external fundamental, respectively. In this sense evidence of FAB in our experiment cannot completely rule out NAB in one-dimensional updating contexts.

Finally, the framework also highlights the important consequences that may result from differences in these biases: namely that with FAB, individuals update in a biased but

[19] Psychologists have also discussed similar potential motivations, see Tetlock and Levi (1982).

consistent manner across both teammates, but with NAB, they update in a biased manner only for themselves, but not for their teammate. The implication is that when taking future decisions involving this fundamental, FAB imposes an additional negative penalty on optimal decision making.

3.5.1 Noisy Attribution Bias

With NAB, individuals may over-attribute positive feedback to their own performance, and/or under-attribute negative feedback to bad luck. Updating consistent with this type of biased updating has been examined in studies with one dimension of uncertainty, for example Eil and Rao (2011), Coutts (2018), Möbius et al. (2014), Buser et al. (2018). Since we consider the additional dimension of uncertainty, ability of their teammate, we must additionally specify that NAB predicts that individuals update using Bayes' rule regarding the performance of their teammate.

Someone who exhibits NAB will update in a way that is consistent with mis-interpreting the strength of the binary signal. That is, when they receive a positive signal, they believe it is more informative about their performance than it really is. When they receive a negative signal, they believe it is less informative about their performance than it really is.

Recall that there are four states of interest: TT, TB, BT, BB. An individual who mis-attributes feedback with respect to noise, is mis-perceiving the informativeness/strength of feedback in at least one of these four states. The most flexible model would allow for mis-perception of strength of feedback potentially all four states. We present a more parsimonious model, which involves mis-perceiving only one of the four states. For our purposes, these models are equivalent.[20]

For parsimony, we thus assume that individuals mis-interpret the strength of the signal when the state is TB. In particular we assume that individuals suffering from NAB believe that the signal is more likely to be positive in state TB than reality. Formally, they believe it has strength $\gamma_p \Phi_{TB}$, where $\gamma_p \geq 1$ in the case of a positive signal, and $\gamma_n \Phi_{TB}$ in the case of a negative signal, where again $\gamma_n \geq 1$. Our specification of the bias is thus similar to the biased updating model of Gervais and Odean (2001).

[20] In particular the structural model we are interested in is responsiveness to feedback on beliefs about performance. For this purpose the model we specify here would be over-identified. While we could take steps to structurally identify mis-perception for each of the four states using our data, we would not put much faith in the resulting parameter estimates. We present this model with an "as-if" interpretation, i.e. we believe this is a parsimonious way to capture self-attribution biases, but we are not arguing this is the mechanical model that subjects have in mind. Moreover, our data suggest that individuals are not accustomed to thinking about the state space in this manner, casting more doubt on such a structural approach.

Biased updating in response to positive and negative feedback through NAB results in more upward biased beliefs about own performance in both cases:

$$\left[b_{t+1}^{1,NAB}|\mathcal{S}=1\right] = \frac{\Phi_{TT}b_t^{TT} + \gamma_p\Phi_{TB}b_t^{TB}}{\Phi_{TT}b_t^{TT} + \gamma_p\Phi_{TB}b_t^{TB} + \Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \geq \left[b_{t+1}^{1,BAYES}|\mathcal{S}=1\right], \tag{7}$$

$$\begin{aligned}\left[b_{t+1}^{1,NAB}|\mathcal{S}=0\right] &= \frac{(1-\Phi_{TT})b_t^{TT} + \gamma_n(1-\Phi_{TB})b_t^{TB}}{(1-\Phi_{TT})b_t^{TT} + \gamma_n(1-\Phi_{TB})b_t^{TB} + (1-\Phi_{BT})b_t^{BT} + (1-\Phi_{BB})b_t^{BB}} \qquad (8)\\ &\geq \left[b_{t+1}^{1,BAYES}|\mathcal{S}=0\right].\end{aligned}$$

With NAB, individuals update asymmetrically about their own performance - they over-weight positive signals and under-weight negative signals relative to a Bayesian. They are perfectly Bayesian with regards to their teammate's performance.

3.5.2 Fundamental Attribution Bias

With FAB, individuals over-attribute positive feedback to their own performance, *at the expense* of the other source of uncertainty, in this case their teammate. Similarly, they under-attribute negative feedback to themselves, and over-attribute it to their teammate. Since experimental research in economics has focused on only one dimension of uncertainty, previous experiments were not able to test for FAB.

FAB takes the same functional form as NAB with regards to own performance.

$$\begin{aligned}\left[b_{t+1}^{1,FAB}|\mathcal{S}=1\right] = \left[b_{t+1}^{1,NAB}|\mathcal{S}=1\right] = \qquad (9)\\ \frac{\Phi_{TT}b_t^{TT} + \gamma_p\Phi_{TB}b_t^{TB}}{\Phi_{TT}b_t^{TT} + \gamma_p\Phi_{TB}b_t^{TB} + \Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \geq \left[b_{t+1}^{1,BAYES}|\mathcal{S}=1\right],\end{aligned}$$

$$\begin{aligned}\left[b_{t+1}^{1,FAB}|\mathcal{S}=0\right] = \left[b_{t+1}^{1,NAB}|\mathcal{S}=0\right] = \qquad (10)\\ \frac{(1-\Phi_{TT})b_t^{TT} + \gamma_n(1-\Phi_{TB})b_t^{TB}}{(1-\Phi_{TT})b_t^{TT} + \gamma_n(1-\Phi_{TB})b_t^{TB} + (1-\Phi_{BT})b_t^{BT} + (1-\Phi_{BB})b_t^{BB}}\\ \geq \left[b_{t+1}^{1,BAYES}|\mathcal{S}=0\right].\end{aligned}$$

The key difference is that with FAB the individual updates in a biased but consistent manner across themselves and their teammate. With NAB, the individual updates in a

biased manner about themselves, but does not inherently care about the assessment of the teammate's performance, and thus updates as a standard Bayesian.

With FAB, individuals update asymmetrically in the *positive* direction regarding their own performance, i.e. they over-weight positive signals, and under-weight negative signals. Regarding their teammate's performance they update asymmetrically in the *negative* direction, i.e. they under-weight positive signals and overestimate negative signals.

$$\begin{aligned}\left[b_{t+1}^{2,FAB}|\mathcal{S}=1\right] &= \frac{\Phi_{TT}b_t^{TT}+\Phi_{BT}b_t^{BT}}{\Phi_{TT}b_t^{TT}+\gamma_p\Phi_{TB}b_t^{TB}+\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}} \\ &\leq \left[b_{t+1}^{2,BAYES}|\mathcal{S}=1\right] = \left[b_{t+1}^{2,NAB}|\mathcal{S}=1\right],\end{aligned} \tag{11}$$

Attribution bias to negative feedback enters as the term $\gamma_n \geq 0$.

$$\begin{aligned}\left[b_{t+1}^{2,FAB}|\mathcal{S}=0\right] &= \frac{(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{BT})b_t^{BT}}{(1-\Phi_{TT})b_t^{TT}+\gamma_n(1-\Phi_{TB})b_t^{TB}+(1-\Phi_{BT})b_t^{BT}+(1-\Phi_{BB})b_t^{BB}} \\ &\leq \left[b_{t+1}^{2,BAYES}|\mathcal{S}=0\right] = \left[b_{t+1}^{2,NAB}|\mathcal{S}=0\right].\end{aligned} \tag{12}$$

4 Experimental Design

4.1 Overview

The experiment was conducted at the WiSo experimental laboratory at the University of Hamburg. A total of 423 students participated in 17 sessions, across two waves in the 2017-18 academic year. Experimental sessions in the first wave lasted approximately 1 hour, and subjects received an average payment of €14. The second wave was identical to the first but had a slight difference in the belief elicitation, and added an additional component where individuals could switch teammates, and hence lasted 1.5 hours with subjects receiving €19.[21]

The experiment consisted of two main parts, following the theory. Part 1 consisted of a 10 minute IQ style test which allowed subjects to be ranked according to their performance. Part 2 consisted of two-person teams, and subjects needed to choose how much to weight the performance of teammate 1 versus teammate 2. Subjects received four rounds of feedback, and took five weighting decisions in total after receiving feedback. A third part, only present in the second wave, presented subjects with the possibility to change

[21] 52% of the participants were Female. 192 subjects participated in the first wave, while 231 subjects participated in the second. Earnings included a €5 show-up fee.

teammates, and consisted of another four rounds of feedback and weighting decisions.

At the beginning of the experiment we provided subjects with the instructions for Part 1 and announced that they would receive the instructions for the other parts as the experiment progressed. In Part 1 subjects had 10 minutes to complete a trivia and logic test consisting of 15 questions. A timer in the upper right corner of the screen continuously informed subjects how much time they had left to finish the test. The instructions stated: "Questions similar to these are often used to measure a person's general intelligence (IQ). Your task is to answer as many of these questions correctly as possible."[22]

Subjects were assigned one of two versions of the test, randomized at the session level, one harder and one easier. This allows us to examine whether the hard-easy effect, see Larrick et al. (2007) and Moore and Small (2007), replicates in our setting, and to examine whether we see differences in belief updating for hard versus easy tests. Subjects were unaware of these differences and were incentivized the same way in both versions: each correct answer would earn 2.5 points while an incorrect answer would be penalized by 1 point. Unanswered questions did not affect the final score. We opted for these incentives to ensure that subjects attempted a question only if they were relatively sure that they knew the answer.[23] Subjects could not score below zero and were paid €0.10 per point earned in Part 1 at the very end of the experiment. At this stage no feedback on performance was given.

Part 2 varied depending on the experimental condition which we manipulated between sessions. The primary treatment manipulation involved whether subjects themselves were members of the team (and hence were reporting beliefs about themselves and their teammate), or whether they were a third party reporting beliefs about a different teammate 1 and teammate 2. We refer respectively to these as Main and Control treatments. In each case payoffs were determined by their reported beliefs and the resulting weighting decision, which follows from the theory. The only difference was that in the Main treatment their own performance was relevant, while in Control, it was another individual's performance. This part is described in more detail below.

Wave 2 differed from Wave 1 primarily in the existence of an additional Part 3, where we first elicited subjects' willingness to pay to switch their teammate. After the elicitation, subjects continued to Part 3 which was the same as Part 2, but with possibly a new

[22]Our priority was in emphasizing the importance of the test to subjects, so that they would care about their ranking. Our questions were gathered from publicly available materials, and while they had a stated use of measuring general intelligence, they cannot be interpreted as true IQ estimates.

[23]This way we insured that the attempted number of questions would carry some informational value, which we use in the later parts of the experiment. If women are more risk averse this could lead to gender differences in number of attempted questions, see Marín and Rosa-García (2011). We do not find any gender differences in number of attempted questions.

teammate. Below we first describe the general setup of Part 2 in the context of the Main treatment. Next we move on to the description of our Control treatment. And finally we explain how we elicited subjects' willingness to pay to switch teammates in Wave 2. Full experimental instructions are presented in the Online Appendix.

4.2 Main Treatment

At the beginning of Part 2, subjects were paired into teams of two which remained constant throughout this part. Their individual performances on the test from Part 1 jointly defined their "team performance" in Part 2. We neither provided subjects with any information about their teammates' identity nor about their teammates' test scores. Subjects were only given information on the number of questions that their teammate *attempted* on the quiz from Part 1. Number attempted does provide some limited information about performance, which generated heterogeneity in prior beliefs which is useful for our analysis involving non-parametric matching on prior beliefs.

We designed the team formation protocol such that both teammates' test scores were compared to the same randomly selected group of 19 other test scores from the experimental session. Each subject could either score in the top 10 (top half) or the bottom 10 (bottom half) of this comparison group of 20, with ties broken randomly. Subjects did not learn their absolute score nor whether they themselves or their teammates belonged to the top or bottom half until the very end of the experiment. Not comparing teammates' scores to each other, but only to the same random comparison group, ensured that the teammates' individual rankings were independent of each other.

Subjects were told that their *individual* financial rewards from Part 2 would depend on their team performance which was determined by the teammates' relative rankings in Part 1 as well as by a "weighting decision" that they would take during Part 2. This weighting decision as well as the relevant performance states follow directly from the theory. The weighting decision only affected subjects' own earnings, and this was emphasized in the instructions. This ensured that social preferences played no role in the weighting decisions.

As in the theory, see Equation 1, payoffs depended on the four relevant performance states, with the positive payment P set to be equal to €10. Thus subjects would earn an amount of €10 (€0) for sure, if both of the teammates were ranked in the top half (bottom half). If one teammate was ranked in the top and the other was ranked in the bottom half, the probability of earning €10 would depend on the weighting decision the took in round

$t, \omega_t \in [0, 1]$.[24] The probability of winning the €10 was given by $\sqrt{\omega_t}$ if a subject scored in the top half (and the teammate in the bottom half) and $\sqrt{1 - \omega_t}$ if a subject scored in the bottom half (and the teammate in the top half).

4.3 Control Treatment

The only difference between the Main and the Control treatment is that in the latter, subjects play the role of a third party who must take the above weighting decision for a team composed of two different individuals. By comparing behavior across the Main and Control treatments, we are able to study any differences that may be present due to ego-relevance of the decision environment, since in the Control treatment, subjects' beliefs and subsequent earnings do not depend on their own performance.

At the beginning of Part 2 in Control, each subject (the "decision maker") was assigned to a team consisting of two randomly selected other subjects (the teammates) from the same session. The decision maker was shown the screenshot of the answers to the IQ quiz of one of the teammates (*teammate 1*) and was provided with information about the number of attempted questions of the other teammate (*teammate 2*). In this way, we ensured that the decision maker in the Control treatment has identical information about all decision-relevant variables as the subjects in the Main treatment (who are themselves in the role of teammate 1). The decision makers' earnings in Part 2 were determined using the analogous rule to the Main treatment, depending on teammates 1 and 2, rather than own performance.

4.4 Belief Elicitation

The main purpose of the weighting decision and its direct relationship with earnings was to provide subjects with a monetary incentive to truthfully report their beliefs about the teammates scoring in the top half. Specifically, subjects were given complete information about the structure of expected payoffs, including the lottery induced by the weight they chose. They were then told that the computer would maximize the probability that they earned the €10 for any given beliefs that they held. After they entered their beliefs, the screen displayed the true (ex-ante) probability of winning the €10 for every possible weight combination. It highlighted the weight that gave them the maximum probability of winning, however they were free to ignore this recommendation and enter any other weight.

[24] We note that subjects saw a transformed weight from 0 to 100 rather than 0 to 1, to make the experiment more intuitive for subjects.

This procedure is novel in its indirect implementation, but shares the same incentive compatibility properties of other probabilistic elicitation procedures such as matching probabilities (Holt and Smith (2009) and Karni (2009)), or the binarized scoring rule (Hossain and Okui (2013)). Our procedure does not require the assumption of risk-neutrality, and requires minimal assumptions of probabilistic sophistication, see Machina (1982). Our method is simple and transparent, and does not suffer from common critiques of these other probabilistic methods, namely that they are difficult for subjects to understand. We show subjects the precise procedure that maps the weight they choose to expected payoffs. Given that this is complicated, we truthfully tell them that when they report their beliefs to us, the computer will calculate for them the probability of earning the €10 for every possible weight between 0 and 1. This is shown to them graphically in z-tree (Fischbacher (2007)), reproduced in Figure 1. At this point the only decision they need to take is choosing the weight that gives them the highest probability of earning the €10.

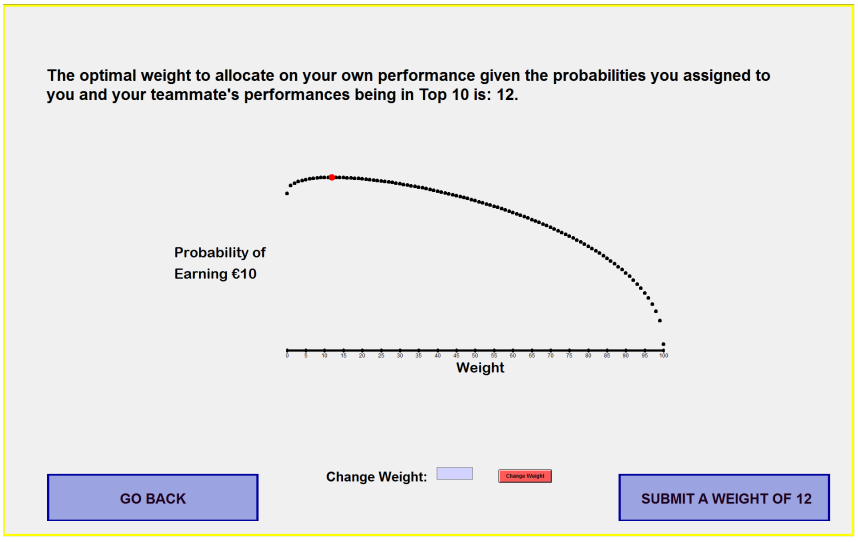


Figure 1: Screenshot of the mapping from chosen weight to probability of winning €10 which was showed to every subject, conditional on the beliefs they entered.

As such, we rely on the incentives to choose optimal weights in order to indirectly incentivize the belief elicitation. So long as individuals prefer to follow the guidance of the weight calculator rather than select a non-recommended weight, this procedure is incentive compatible. Note that it is also true that if subjects choose to enter different weights from those suggested, we are no longer able to claim incentive compatibility. Reassuringly, only 11% of weights did not correspond to the suggested optimal.[25]

[25]Results are not substantively affected excluding these observations. Note that theoretically there are

For each elicitation in Wave 1 subjects entered beliefs for the probability that teammate 1 scored in the top half, and the probability that teammate 2 scored in the top half. In Wave 2 beliefs were *additionally* elicited about the probabilities of all four possible states of TT, TB, BT, and BB. Screenshots of the procedure can be seen in the instructions in the Online Appendix.

Without additional assumptions, calculating the optimal weight requires knowledge of the four relevant states (TT, TB, BT, BB), as beliefs are not independent in later rounds, as noted in the theory section. In Wave 1 we assumed independence between beliefs about performance of the teammates, in order to calculate the probabilities of the four states. In Wave 2 we elicited the identical probabilities as in Wave 1, calculated state-wise probabilities assuming independence, but then we gave subjects full freedom to re-allocate these probabilities to the four relevant states as they saw fit. Reassuringly, 90% of the time subjects also chose to not alter beliefs in the four states, that is they also followed the independence assumption.[26] In the Online Appendix we also show beliefs are nearly identical across the two waves.

4.5 Feedback

Once their weight was submitted subjects received feedback from a "Team Evaluator", represented as a cartoon figure. Binary signals were framed as positive or negative team feedback, corresponding to the Team Evaluator giving a "Green Check" or "Red X" respectively. If both teammates scored in the top half, the Team Evaluator gave a Green Check with $\Phi_{TT} = 90\%$ and a Red X with 10% probability. If one teammate scored in the top half and the other scored in the bottom half, then the Team Evaluator gave a Green Check or a Red X with $\Phi_{TB} = \Phi_{BT} = 50\%$ probability. If both teammates scored in the bottom half, then the Team Evaluator would give the Red X with 90% and a Green Check with $\Phi_{BB} = 10\%$ probability. Note that the feedback received from the Team Evaluator was related to the actual performance of the teammates in Part 1 of the experiment and did not depend on the previous beliefs reported by subjects nor the previous weights submitted. This ensured that subjects did not have incentives to "experiment" with their chosen

different combinations of beliefs (in particular sharing the same ratio) that lead to the same optimal weight. In theory it is thus possible that subjects are able to arrive at the optimal weight, but intentionally try to deceive the experimenters by reporting a different combination that leads to their preferred weight. We do not find this likely.

[26] For the 10% that did not, the average difference in the belief reported was less than one percentage point. Results are robust to excluding these observations. Piloting suggested it was very unnatural for subjects to initially think about the probabilities of these four states. For this reason we felt it was important to first ask about the probability of teammate 1 and 2 being in the top half, then calculate the probability of these four states.

beliefs and weights to learn more about their rankings.

After receiving the Team Evaluator's feedback, subjects entered the next elicitation stage where they had to again report their beliefs that the teammates scored in the top half. Subsequently, the computer gave them a new weight recommendation which they could review and submit. This process was repeated four times. That is in total, subjects reported their beliefs about their and their teammate's performance and submitted a weight five times and received feedback from a Team Evaluator four times. At the beginning of the Part 2, subjects were told that one of the five weighting decisions would be selected at random and the probability of winning the €10 would depend on the weighting decision as well as on the teammate's performance as explained above.[27]

4.6 WTP

At the end of Part 2, we asked subjects their maximum willingness to pay (WTP) to switch their teammates for Part 3, i.e. be randomly rematched with a new teammate. Part 3 was otherwise identical to Part 2. We elicited WTP using the BDM mechanism of Becker et al. (1964). The mechanism asked subjects to enter any amount between €0 and €5 as their maximum willingness to pay. The lottery would then choose a number in [€0, €5] interval and subjects would switch their teammates if their maximum WTP was above the chosen number and keep their teammates if this maximum is below that number.

5 Hypotheses

We present our hypotheses in pairs, which will respectively refer to the Bayesian benchmark, and the benchmark generated by our Control treatment.

5.1 Belief Formation

Regarding the Bayesian benchmark for belief formation, it is perfectly admissible for individual agents to forecast their ability with error (e.g. overconfidence or underconfidence), but these errors should be mean zero on average. Following, Dubra and Benoît (2011),

[27] For more discussion on incentive compatibility of paying for one randomly selected decision in experiments see Azrieli et al. (2018). Note that in Wave 2 there is an additional paid Part 3, however subjects are not aware of its structure until completing Part 2. Before the actual commencement of Part 2, subjects had to answer five control questions that were aimed at ensuring subjects' understanding of the payment calculation, Team Evaluator's feedback, and the weighting function. Subjects were only allowed to start Part 2 of the experiment and enter their first belief when the experimenter had checked that the answered provided were correct.

in this benchmark case we require that beliefs of scoring in the top half (the top 50%) are on average equal to 0.5. As Dubra and Benoît (2011) demonstrate in their Theorem 3, if this does not hold true in the population, then such beliefs cannot be rationalized. Rationalized is used in the sense that beliefs have been formed and updated according to the properties of Bayes' rule.[28]

Our first null hypothesis of interest concerns whether there is overconfidence in the Main treatment. Let $b_0^{1,M}$ be the average initial ($t = 0$) belief about one's own probability of scoring in the top half, where the superscript M stands for Main treatment and 1 indicates that it is teammate 1.

Hypothesis 1:

$$b_0^{1,M} = 0.5.$$

If $b_0^{1,M}$ were significantly greater than 0.5, this would suggest the presence of overconfidence, while the opposite would suggest underconfidence.

However, we also can provide more stringent tests of overconfidence, using our Control treatment. The reason is as noted above, finding evidence of over or underconfidence, may reflect the presence of other irrationalities in belief formation process. Hence we present the paired null hypothesis:

Hypothesis 1*:

$$b_0^{1,M} = b_0^{1,C},$$

where C indicates the Control treatment. We believe that examining these two hypotheses together presents a much stricter test for over or underconfidence.

5.2 Belief Updating

Continuing with our first benchmark, we take Bayes' rule as the initial benchmark for the analysis of belief updating. However, previous studies (Grether, 1980, 1992; Holt and Smith, 2009; Möbius et al., 2014; Coutts, 2018; Buser et al., 2018) have shown that there

[28] As Dubra and Benoît (2011) note, it is possible that individuals are not overconfident, but do form irrational beliefs, and generate patterns that happen to be consistent with the presence of overconfidence.

are important deviations from Bayes' rule in belief updating. Due to expected deviations from Bayes' rule, it is important to define a further control group to provide the relevant counterfactual updating patterns necessary to establish any deviations in behavior. Thus, we additionally make comparisons between the Main and Control treatments of the experiment.

Recall that the signal strengths are given by $\Phi_{TT} = 0.9$, $\Phi_{TB} = \Phi_{BT} = 0.5$, and $\Phi_{BB} = 0.1$. Hence, TB and BT contain the same likelihood ratios of observing positive relative to negative signals, 1. The implication is that *the optimal weight will be constant*: that is, it will be non-responsive to feedback in our context. As delegation weights are a function of beliefs, we focus our hypotheses on beliefs themselves. Bayesian updating for teammate 1 follows directly from Equations 5 and 6:

$$\begin{aligned} \left[b_{t+1}^{1,BAYES} | + \right] &= \frac{0.9b_t^{TT} + 0.5b_t^{TB}}{0.9b_t^{TT} + 0.5b_t^{TB} + 0.5b_t^{BT} + 0.1b_t^{BB}} \\ \left[b_{t+1}^{1,BAYES} | - \right] &= \frac{0.1b_t^{TT} + 0.5b_t^{TB}}{0.1b_t^{TT} + 0.5b_t^{TB} + 0.5b_t^{BT} + 0.9b_t^{BB}}. \end{aligned} \tag{13}$$

Analogously for teammate 2:

$$\begin{aligned} \left[b_{t+1}^{2,BAYES} | + \right] &= \frac{0.9b_t^{TT} + 0.5b_t^{BT}}{0.9b_t^{TT} + 0.5b_t^{TB} + 0.5b_t^{BT} + 0.1b_t^{BB}} \\ \left[b_{t+1}^{2,BAYES} | - \right] &= \frac{0.1b_t^{TT} + 0.5b_t^{BT}}{0.1b_t^{TT} + 0.5b_t^{TB} + 0.5b_t^{BT} + 0.9b_t^{BB}}. \end{aligned} \tag{14}$$

5.2.1 Empirical Updating Framework

Here we examine the implications of the theory for the empirical framework, which follows Grether (1980) and Möbius et al. (2014). Bayes' rule can be written in the following form, considering binary signals, $s_t = k \in \{0, 1\}$, and letting $\hat{b}_t^i$ be the belief at time t of the subject about teammate $i \in \{1, 2\}$:

$$\frac{\hat{b}_t^i}{1 - \hat{b}_t^i} = \frac{\hat{b}_{t-1}^i}{1 - \hat{b}_{t-1}^i} \cdot LR_t^i(k) \tag{15}$$

where $LR_t^i(k)$ is the likelihood ratio of observing signal $s_t = k \in \{0, 1\}$ when updating beliefs about teammate i. For the sake of clarity, we focus this discussion from the perspective of updating beliefs about teammate 1; results for teammate 2 are analogous. From the theory which includes potential attribution biases, the perceived likelihood ratio

of observing a positive signal conditional on teammate 1 being in the top half is:

$$\frac{0.9b_t^{TT} + \gamma_p 0.5b_t^{TB}}{b_t^{TT} + b_t^{TB}},$$

where $\gamma_p = 1$ indicates the likelihood ratio a Bayesian perceives. The perceived likelihood of observing a positive signal conditional on teammate 1 being in the bottom half is:

$$\frac{0.5b_t^{BT} + 0.1b_t^{BB}}{b_t^{BT} + b_t^{BB}}$$

Recalling that $b_t^1 = b_t^{TT} + b_t^{TB}$, the ratio, $\hat{LR}_t^1(1)$, is thus:

$$\frac{0.9b_t^{TT} + \gamma_p 0.5b_t^{TB}}{0.5b_t^{BT} + 0.1b_t^{BB}} \cdot \frac{1 - b_t^1}{b_t^1} \geq 1$$

Similarly, the ratio, $\hat{LR}_t^1(0)$, is:

$$\frac{0.1b_t^{TT} + \gamma_n 0.5b_t^{TB}}{0.5b_t^{BT} + 0.9b_t^{BB}} \cdot \frac{1 - b_t^1}{b_t^1} \leq 1$$

Taking natural logarithms of both sides of Equation 15 and using an indicator function, $I\{s_t = k\}$, for the type of signal observed,

$$logit(\hat{b}_t^i) = logit(\hat{b}_{t-1}^i) + I\{s_t = 1\} \ln\left(\hat{LR}_t^i(1)\right) + I\{s_t = 0\} \ln\left(\hat{LR}_t^i(0)\right). \quad (16)$$

The empirical model nests this Bayesian benchmark as follows,

$$logit(\hat{b}_{jt}^i) = \delta logit(\hat{b}_{j,t-1}^i) + \beta_1 I(s_{jt} = 1) \ln\left(\hat{LR}_t^i(1)\right) + \beta_0 I(s_{jt} = 0) \ln\left(\hat{LR}_t^i(0)\right) + \epsilon_{jt}. \quad (17)$$

δ captures the weight placed on the log prior odds ratio. β_0 and β_1 capture responsiveness to either negative or affirmative signals respectively. In the context of the experiment, $s_{jt} = 1$ corresponds to a positive signal, while $s_{jt} = 0$ corresponds to a negative signal. Since $I(s_{jt} = 0) + I(s_{jt} = 1) = 1$ there is no constant term. ϵ_{jt} captures non-systematic errors, noting the use of j to identify the experimental subject.

Bayes' rule is a special case of this model when $\delta = \beta_0 = \beta_1 = 1$. Let superscripts M

and C denote these coefficients on the Main versus Control treatments. $\delta^{1,M}$ will be used to describe the coefficient of δ for teammate 1 in the main sessions (i.e. the individual themselves), $\delta^{2,M}$ describes the coefficient of δ for teammate 2 in the main sessions. Similarly for control (C), with analogous definitions for β_1 and β_0.

What are the implications of NAB and FAB for the framework? First note that $\hat{LR}_t^1(1) \geq LR_t^1(1)$ and $\hat{LR}_t^1(0) \geq LR_t^1(0)$. Bayesian posteriors result in a weight of $\beta_1 = 1$ or $\beta_0 = 1$ on $LR_t^1(1)$ or $LR_t^1(0)$ respectively. For an individual suffering from FAB or NAB who perceives greater likelihood ratios, estimates of β_1 will be biased upwards for teammate 1, while estimates of β_0 will be biased downwards for teammate 1.[29] In words, after a positive signal, someone with NAB or FAB believes that the signal was more indicative of being in the top than it really is. After a negative signal they believe that the signal is less indicative about being in the bottom. Note that for teammate 2, with NAB there are no distortions in updating, while with FAB the distortions are opposite those of teammate 1 (negative asymmetry). The remaining null hypotheses are as follows.

Hypothesis 2:

Belief updating follows the mechanics of Bayes' rule:
$\delta^M = 1;\ \ \beta_1^M = 1;\ \ \beta_0^M = 1$

Hypothesis 2*:

Belief updating is the same across Main and Control treatments:
$\delta^M = \delta^C;\ \ \beta_1^M = \beta_1^C;\ \ \beta_0^M = \beta_0^C$

Hypothesis 3:

Noisy Attribution Bias (Bayesian benchmark):
$\delta^{1,M} = 1\ ;\quad \beta_1^{1,M} > 1;\ \ \beta_0^{1,M} < 1$
$\delta^{2,M} = 1\ ;\quad \beta_1^{2,M} = 1;\ \ \beta_0^{2,M} = 1$

[29] That β_1 is biased upwards is straightforward, since $ln(\hat{LR}_t^1(1)) \geq 0$ so a Bayesian response to in $\hat{LR}_t^1(1)$ will manifest itself as an over-response to the smaller unbiased $LR_t^1(1)$. β_0 is biased downwards because $ln(\hat{LR}_t^1(0)) \leq 0$ so a Bayesian response to in $\hat{LR}_t^1(0)$ will manifest itself as an under-response to the smaller (more negative, i.e. larger in absolute value) $LR_t^1(0)$.

Hypothesis 3*:

Noisy Attribution Bias (Control benchmark):
$\delta^{1,M} = \delta^{1,C}; \quad \beta_1^{1,M} > \beta_1^{1,C}; \;\; \beta_0^{1,M} < \beta_0^{1,C}$
$\delta^{2,M} = \delta^{2,C}; \quad \beta_1^{2,M} = \beta_1^{2,C}; \;\; \beta_0^{2,M} = \beta_0^{2,C}$

Hypothesis 4:

Fundamental Attribution Bias (Bayesian benchmark):
$\delta^{1,M} = 1\ ; \quad \beta_1^{1,M} > 1; \;\; \beta_0^{1,M} < 1$
$\delta^{2,M} = 1\ ; \quad \beta_1^{2,M} < 1; \;\; \beta_0^{2,M} > 1$

Hypothesis 4*:

Fundamental Attribution Bias (Control benchmark):
$\delta^{1,M} = \delta^{1,C}\ ; \quad \beta_1^{1,M} > \beta_1^{1,C}; \;\; \beta_0^{1,M} < \beta_0^{1,C}$
$\delta^{2,M} = \delta^{2,C}\ ; \quad \beta_1^{2,M} < \beta_1^{2,C}; \;\; \beta_0^{2,M} > \beta_0^{2,C}$

The hypothesis of NAB states that individuals will update asymmetrically in the positive direction about their own performance relative to either the Bayesian prediction or the control, but will update identically about a teammate relative to these benchmarks. FAB states that individuals will exhibit positive asymmetry regarding the own performance (as in NAB), but will exhibit negative asymmetry for their teammate's performance, in both cases relative to the respective benchmarks.

6 Results

6.1 Prior Beliefs

First we investigate whether prior beliefs are well calibrated in the first round. Following Hypothesis 1 we examine the Main treatment, where individuals were in the position of teammate 1, and thus were estimating beliefs about their own performance. In the Main treatment the average reported belief about being in the top 50% (top half) is 66.4%, presenting very suggestive evidence of significant overconfidence. This is confirmed by a Wilcoxon rank-sum test, which rejects that this is equal to 50% at the 1% level. Hence

this data cannot be rationalized using the test of Dubra and Benoît (2011), and appear to exhibit overconfidence, evidence against Hypothesis 1.

We can also can examine average reported beliefs for those in the Control treatment, who estimated the performance of another, randomly selected individual who was in the position of teammate 1. The average reported belief for this individual being in the top 50% was 56.3% which is also significantly different from 50% at the 1% level using a Wilcoxon signed rank test. Hence we have evidence that these beliefs also "appear" overconfident. Since this does not reflect overconfidence in the traditional sense, as it does not involve estimation of one's own performance, but rather of another's performance, this finding is surprising.[30]

Figure 2 presents the distribution of beliefs about teammate 1 by treatment. A Kolmogorov-Smirnov test confirms what can be seen visually, that the distribution of beliefs for teammate 1 across the two treatments is significantly different at the 1% level. Median beliefs are skewed upwards, 59.5% in control, and 71.5% in main, remarkably the interquartile range for beliefs about own performance (Main treatment) spans 50%-90%. All of these results point to very high degrees of overconfidence, yet it is important to account for the fact that similar, though greatly muted, patterns are observed in the Control.

Thus, following Hypothesis 1*, we compare beliefs across the two treatments, Main and Control. In fact we can reject equality of mean prior beliefs across the two settings at the 1% level (Wilcoxon rank-sum test p-value: 0.0005). This provides robust evidence that what we are observing in the Main treatment does reflect true overconfidence.

Regarding beliefs about teammate 2, these do not exhibit any differences between Main and Control, respectively the belief that teammate 2 is in the top 50% is 53.4% and 54.3%, not statistically different from one-another.[31] Figure 3 presents analogous distributions for teammate 2. While there appear to be some differences in distribution, e.g. less dispersion in Main, a Kolmogorov-Smirnov test cannot reject equality of the distributions (p-value: 0.289).

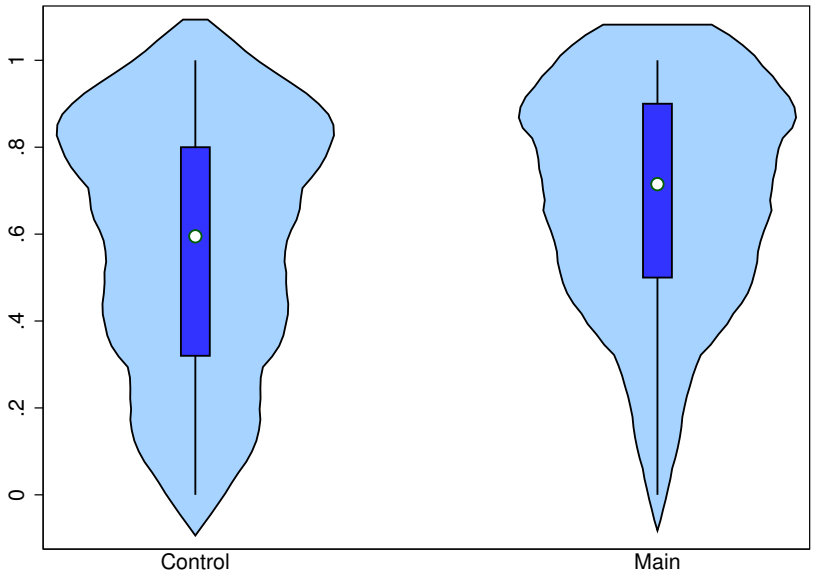
We also note that our hard-easy manipulation was successful. More details are provided in the Online Appendix, however individuals rate themselves in the top half with

[30] One potential explanation is that overconfidence enters through one's estimations of others' performance. Since individuals were able to observe teammate 1's answers, they may make comparisons between their own answers, and that of teammate 1, which could lead to "spillovers" from overconfidence. Contrary to this potential explanation is that beliefs about teammate 1 in Control are not significantly greater than beliefs about teammate 2, for which observing answers was not possible.

[31] The Wilcoxon rank-sum p-value is 0.5723. Additionally a Wilcoxon signed rank test also reject the hypotheses that these beliefs are equal to 50% at the 1% level, again pointing to an upward bias in beliefs that cannot be explained by overconfidence.

72% probability when the test was easy, and 62% when the test was hard.[32] Regarding gender, we find evidence that men are more overconfident than women. We provide further details in the Online Appendix, now we defer discussion about gender differences to results on belief updating, where we can better control for subject priors.[33]

Figure 2: Distribution of Beliefs about Teammate 1 by Treatment

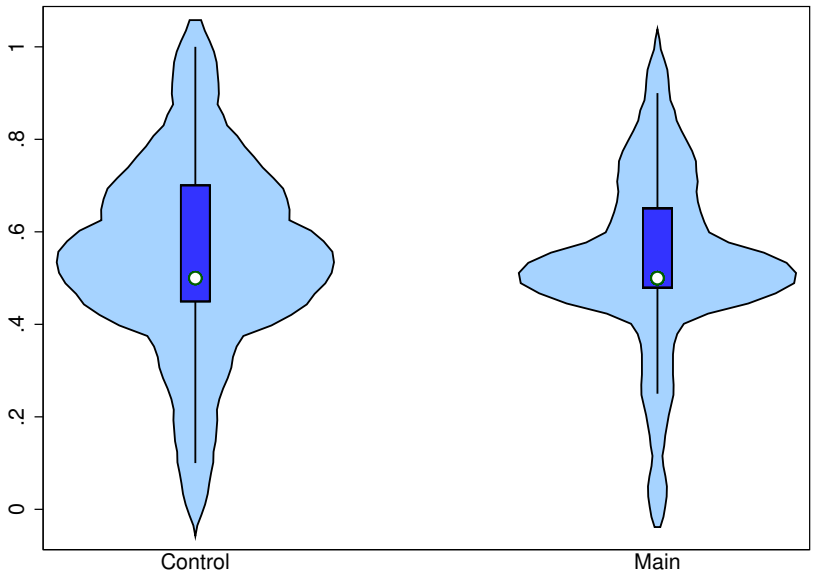


Violin plot of the distribution of beliefs, median, and interquartile range. Main: Belief about own performance. Control: Beliefs about other Teammate 1's performance.

[32] We also find the hard-easy effect in the Control treatment: 61% of our subjects ranked teammate 1 in the top half when the test was easy, compared to 52% when the test was hard.

[33] More details about overconfidence and updating by gender are available in the Online Appendix. There are gender differences in performance driven by poor scores at the lower level, and if we consider only the top 2/3 of scores, these differences disappear. For the analysis in the Online Appendix, performance is controlled for.

Figure 3: Distribution of Beliefs about Teammate 2 by Treatment



Violin plot of the distribution of beliefs, median, and interquartile range. Main: Individual is on the team. Control: Individual is not on the team.

6.2 Belief Updating

To distinguish the types of self-attribution bias we discuss in our theory, we require a structural model of belief updating for our primary empirical analysis. Later, we investigate updating biases taking a non-parametric approach, free of structural assumptions. This allows us to statistically distinguish posteriors in Main versus Control, accounting for differences in initial priors. While our main focus is on beliefs, we also present an analysis of chosen weights in Appendix A as well as WTP to be matched to a new teammate 2 in Appendix B.[34]

6.2.1 Structural Framework

The primary analysis follows the framework outlined in Section 5.2.1. Table 1 presents the main specification for updating of beliefs about teammate 1 for the Main and Control treatments. Recall that in the Main treatment, subjects update about their *own* performance, while in the Control, they act as a third party who is choosing for a team of two other individuals, and thus are updating beliefs about teammate 1 of that team. Our sam-

[34]It is important to remember that chosen weights follow directly from beliefs. While there are some interesting patterns in WTP, these as well partly reflect differences in beliefs.

ple includes all updates from both waves, in Part 2 and 3.[35]

Table 1: Updating Beliefs about Teammate 1

Regressor	(1) Main Treatment	(2) Control Treatment
δ	0.734^{***}	0.751^{***}
	(0.054)	(0.045)
β_1	0.573^{***}	0.506^{***}
	(0.071)	(0.075)
β_0	0.260^{***}	0.507^{***}
	(0.060)	(0.061)
P-Value ($\delta = 1$)	0.0000	0.0000
P-Value ($\beta_1 = 1$)	0.0000	0.0000
P-Value ($\beta_0 = 1$)	0.0000	0.0000
P-Value ($\beta_1 = \beta_0$)	0.0038	0.9906
R^2	0.56	0.60
Observations	863	829
P-Value [Chow-test] for δ (Regressions (1) and (2))		0.8089
P-Value [Chow-test] for β_1 (Regressions (1) and (2))		0.5152
P-Value [Chow-test] for β_0 (Regressions (1) and (2))		0.0040
P-Value [Chow-test] for $(\beta_1 - \beta_0)$ (Regressions (1) and (2))		0.0231

Analysis uses OLS regression. Difference is *significant from 1* at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level. R^2 corrected for no-constant. δ is the coefficient on the log prior odds ratio. β_1 and β_0 are coefficients on the log likelihood of observing positive and negative signals respectively. Constant omitted because of collinearity. Bayesian updating corresponds to $\delta = \beta_1 = \beta_0 = 1$. $\beta_1, \beta_0 < 1$ indicates conservative updating. $\beta_1 - \beta_0 > 0$ indicates positive asymmetric updating.

Our first observation is that that Hypothesis 2 is rejected: updating is not Bayesian in the Main treatment, and this additionally is true for the Control. This can be seen as all coefficients are significantly different from the Bayesian prediction of 1, indicated by asterisks in Table 1. From now on, we thus focus on only our hypotheses which compare updating to Control, rather than the Bayesian benchmark.

From Table 1 Column 1 one can see that positive signals are given significantly more weight than negative signals (positive asymmetry), when updating is about one's own per-

[35]Other samples excluding Part 3 are presented in the Online Appendix, with similar results. We follow sampling restrictions which have been common in the literature: excluding boundary observations and excluding updates in the wrong direction. Because we deal with two dimensions of uncertainty, we define an update as being in the wrong direction if a subject updates at least one belief in the wrong direction, without compensating by adjusting the other belief in the correct direction. We provide more details on this in our discussion on sampling restrictions in the Online Appendix.

formance. The positive asymmetry observed is significant at the 1% level. No asymmetry is observed in column 2, in the Control treatment, for updating about another individual's performance.

Thus, Hypothesis 2* is rejected, updating is not the same across the Main and Control treatments - notably $\beta_0^{1,M} < \beta_0^{1,C}$. While $\beta_1^{1,M} > \beta_1^{1,C}$, this is not statistically significant. However, importantly, $\beta_1^{1,M} - \beta_0^{1,M} > \beta_1^{1,C} - \beta_0^{1,C}$, is significantly different from 0 at the 5% level, indicating that individuals exhibit more (positive) asymmetry in updating in Main than in Control.

Are these patterns consistent with the two types of attribution bias outlined in Hypotheses 3* and 4*? In fact, positive asymmetry is predicted by both NAB and FAB. In order to distinguish them, we need to additionally examine updating about teammate 2. NAB posits that updating should not be biased about teammate 2, i.e. that individuals mis-attribute feedback about their own performance, relegating the difference to noise, but not to their teammate. FAB on the other hand, posits that individuals mis-attribute feedback about their own performance specifically with regard to their teammate. FAB predicts that updating about teammate 2 will thus be negatively asymmetric, over-weighting negative relative to positive signals.

Table 2: Updating Beliefs about Teammate 2

Regressor	(1) Main Treatment	(2) Control Treatment
δ	0.770^{***}	0.717^{***}
	(0.048)	(0.050)
β_1	0.398^{***}	0.491^{***}
	(0.056)	(0.070)
β_0	0.248^{***}	0.418^{***}
	(0.043)	(0.061)
P-Value ($\delta = 1$)	0.0000	0.0000
P-Value ($\beta_1 = 1$)	0.0000	0.0000
P-Value ($\beta_0 = 1$)	0.0000	0.0000
P-Value ($\beta_1 = \beta_0$)	0.0358	0.3708
R^2	0.53	0.50
Observations	1016	916
P-Value [Chow-test] for δ (Regressions (1) and (2))		0.4408
P-Value [Chow-test] for β_1 (Regressions (1) and (2))		0.2977
P-Value [Chow-test] for β_0 (Regressions (1) and (2))		0.0235
P-Value [Chow-test] for $(\beta_1 - \beta_0)$ (Regressions (1) and (2))		0.4728

Analysis uses OLS regression. Difference is *significant from 1* at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level. R^2 corrected for no-constant. δ is the coefficient on the log prior odds ratio. β_1 and β_0 are coefficients on the log likelihood of observing positive and negative signals respectively. Constant omitted because of collinearity. Bayesian updating corresponds to $\delta = \beta_1 = \beta_0 = 1$. $\beta_1, \beta_0 < 1$ indicates conservative updating. $\beta_1 - \beta_0 > 0$ indicates positive asymmetric updating.

Table 2 presents the analogous regressions for teammate 2 in Main (column 1) and Control (column 2). Interestingly, patterns are similar to updating for teammate 1, though less pronounced. In fact there is evidence of positive asymmetry for teammate 2 in the main treatment, significant at the 5% level. One can reject that the coefficient β_0 is the same across the Main and Control at the 5% level. Again with respect to the hypotheses of interest, Bayesian updating is rejected, as the coefficients differ significantly from 1. Similarly the hypothesis of equivalent updating across the Main and Control treatments (Hypothesis 2*) is rejected.

As column 1 in Table 2 demonstrates, subjects exhibit positive asymmetry for teammate 2 in the Main treatment, when they themselves are a member of the team. This asymmetry is significant at the 5% level, which is not consistent either with NAB or

FAB.[36] Thus, we find evidence of positive asymmetry both for teammate 1 (self) and teammate 2 when one is a member of the team (Main treatment), but no significant asymmetry when one is not a member of the team (Control treatment). The strongest finding is that of under-weighting negative feedback in the Main treatment. This is significantly lower both for teammate 1 and 2, relative to the Control treatment.

It is important to note that while these patterns do not support the predictions of NAB and FAB, there is significant mis-attribution in the Main treatment. When receiving positive signals, individuals are over-attributing positive feedback to themselves, and there are some suggestive patterns that they are under-attributing positive feedback to their teammate, as β_1 in Table 1 is significantly greater than β_1 in Table 2 at the 1% level (Chow test p-value 0.0068). When receiving negative signals, they are greatly over-attributing these to noise, relative to how individuals update in the control treatment. Such behavior could indicate the use of a more blunt form of noisy attribution bias in which individuals under-weight negative feedback without regard to other sources of uncertainty. Importantly, this provides direct evidence that individuals' biased attribution patterns can affect not just their own beliefs but also their assessments of others.[37]

Finally we discuss briefly differences in updating as a result of hard-easy test differences, and additionally by gender. The Online Appendix presents these tables. There is some evidence that there is greater asymmetry when the test is hard rather than easy, though these differences are not significant. Interestingly, nearly all of the asymmetry observed appears to be generated by men. That is, we find similar patterns when examining the results for males only, but find no evidence of asymmetric updating for females only. These patterns are true both for updating about teammate 1 and teammate 2.

6.2.2 Non-Parametric Inference

While the previous section provides significant evidence of differential updating between the Main and Control treatments, one can also examine whether updating is different between these treatments outside of the structural framework.

As a first look we examine how prior beliefs evolve after feedback. We do this in two ways, the first way is to present the raw data on beliefs after each round of feedback,

[36] Note that the level of magnitude of asymmetry is not significantly different in Main versus Control, as there is also slight positive asymmetry in the Control treatment. However, the weight on the response to negative signals β_0 is significantly different across the two treatments at the 5% level.

[37] It could suggest that individuals treat the team as one unit - which they then update according to the NAB model. However, patterns which could be interpreted as evidence against this hypothesis is that (i) we do not see strong evidence of overconfidence regarding teammate 2's performance in the Main treatment relative to the Control, and (ii) there is no such asymmetry in the Control treatment, indicating that individuals are not biased when it comes to updating about a team (in general).

relative to the Bayesian benchmark. For this we consider rounds 1 through 5 in Part 2 common to all subjects, but not Part 3.[38] The second way is to present a flexible polynomial smoothing plot of the relationship between priors and posteriors, pooling all of the updating rounds. In this second case we separately conduct this exercise for positive versus negative feedback. We include data from Parts 2 and 3, and we follow the same sampling restrictions as Tables 1 and 2 for comparability.[39]

For this first look, in Appendix Figures C.1 and C.2 we plot beliefs by round of teammate 1 and teammate 2 respectively, showing both Main versus Control, as well as the Bayesian predictions. Average posteriors at the end of Part 2 are 65.6% in the Main treatment compared to 52.9% in the Control, a difference of 12.7 percentage points. While significantly different, this is clearly not an adequate comparison to detect differences in updating patterns, due to differences in prior beliefs (66.4% and 56.3% respectively). Of note is that beliefs in the Control appear to decrease more than those in the Main treatment, leading to greater deviations from the Bayesian predictions in the Main treatment, both for teammate 1 and 2. Figure C.3 more specifically presents the deviation from the Bayesian posterior at the end of Part 2, showing that there are suggestive differences in the Main treatment, but not in Control.[40]

For the second look, Figures C.4 through C.7 show the relationship between priors and posteriors, separating Main and Control for positive and negative signals, as well as for teammate 1 and 2 respectively. The first pattern is that updating appears "flatter" than what Bayes' rule prescribes. In the empirical framework, this is captured by the coefficient $\delta < 1$, which indicated that subjects were updating as if priors were weighted more towards 50%. Note also that one can see the substantial conservatism, as subjects do not on average update sufficiently in the direction feedback indicates, relative to the Bayesian prediction. Beyond this, one can see that for teammate 1, conditional on the same priors, posteriors in Main show an upward bias relative to Control, for both positive and negative signals, consistent with the patterns in Table 1. For teammate 2 similar biased patterns are visible only for responses to negative signals.

Overall it appears that the biased updating patterns revealed in the previous section translate to differences in posterior beliefs, relative to the predictions of Bayes' rule. However it is critical to control for the differences in prior beliefs across Main and Con-

[38] We do not present Part 3 due to the significantly smaller sample, as a result of the exclusion of Wave 1 as well as subjects re-matched to new teammates in Wave 2.

[39] We could also present average posteriors conditional on the valence of the signal. However this exercise is confounded by the fact that negative signals are correlated with priors. Our main non-parametric analysis uses matching which will control for prior beliefs.

[40] These differences are not always significant at conventional levels, and so are only suggestive. See the Appendix for more details on the statistical tests.

trol (particularly for teammate 1). The next subsection presents a matching strategy which will control for initial priors in detecting differences in updating across Main and Control.

6.2.3 Matching on Priors

While there is evidence that beliefs are updated differently in the Main versus Control relative to the Bayesian benchmark, it is also important to examine the extent to which updating differs across the Main and Control without relying on the Bayesian benchmark or a quasi-Bayesian framework. In this subsection we present a non-parametric analysis of updated beliefs, which utilizes a matching strategy that matches the Main and Control subjects on their prior beliefs, and then compares their posteriors at the end of Part 2 after four rounds of feedback.[41] By matching on prior beliefs we are able to step away from the reliance on Bayes' rule, and instead ask the following question. Given the same priors, do subjects arrive at different posteriors about their own abilities (Main treatment) versus the abilities of a stranger (Control treatment)? Beyond this, to ensure that these matched subjects face the same number of positive and negative signals, we force exact matching on the number of negative signals received.

Table 3 presents the results of this exercise reporting average treatment effects (ATE). In fact, the matching strategy reveals that individuals who are updating about their own performance (Main treatment) end up with posteriors that are 6.6 to 8.4 percentage points greater than those updating about the performance of a stranger, conditional on having the same priors and facing the same sequence of signals.[42] This indicates that information processing differs across these two groups. Regarding teammate 2, the analogous difference between the Main and Control is 4.5 percentage points, which is not significant at conventional levels.

The empirical framework suggests this difference in updating is driven primarily by under-responsiveness to negative signals. To investigate this, Table 4 presents matching estimates for each of the possible sequences of signals observed separately. Consistent with the structural framework, receiving 4 negative signals (0 positive) turns out to reveal the greatest bias between Main versus Control: subjects with the same priors end up an estimated 18.5 percentage points more confident when they are estimating their own performance. The only other significant effect is an equally balanced sequence of 2 positive

[41] Since we are working with final posteriors, Part 3 is not feasible since it was not included in Wave 1, and additionally involves some re-matching of teammates, invalidating these posteriors for this purpose.

[42] We set a caliper of 0.03, meaning that priors of matched neighbors must be within 3 percentage points. There is a trade-off: too small and we are unable to generate sufficient matches, too large and we reduce the quality of the matches.

Table 3: Main vs Control: Belief Teammate 1 Top

	(1) 1 Neighbor	(2) 2 Neighbors
ATE	0.084^{***} (0.032)	0.066^{**} (0.029)
Observations	372	372

Analysis uses nearest neighbor matching, with replacement when > 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. All matches received the exact same distribution of signals.

Table 4: Main vs Control: Belief Teammate 1 Top by Distibution of Received Signals

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 4 −
ATE	-0.016 (0.073)	0.105 (0.083)	0.134^{***} (0.047)	-0.025 (0.087)	0.185^{**} (0.084)
Observations	72	68	100	60	72

Analysis uses nearest neighbor matching with 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. Each column restricts sample to specific distribution of negative signals received (out of 4 total signals).

and 2 negative signals. While this is supportive of the structural results, the effect is not monotonic in the number of negative signals.[43]

Regarding the non-parametric estimates of the effect of differential updating about teammate 2 when one is a member of the team (Main treatment) versus not (Control), analogous regressions are presented in Tables 5 and 6. The estimated ATE is between 4.0 and 4.5 percentage points greater posterior belief about one's teammate in Main relative to Control, however this is not statistically significant at conventional levels. Of note is that when examining separately the ATE estimates for different distributions of negative signals received, receiving all negative signals is associated with a large and significant effect. Individuals with the same priors about teammate 2 in Main and Control who receive only negative signals end up with posteriors about teammate 2 that are 14 percentage

[43] Additionally the estimates are noisier than the overall matching estimates due to smaller sub-samples.

points greater in Main relative to Control.

Table 5: Main vs Control: Belief Teammate 2 Top

	(1) 1 Neighbor	(2) 2 Neighbors
ATE	0.045 (0.035)	0.040 (0.030)
Observations	394	394

Analysis uses nearest neighbor matching, with replacement when > 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. All matches received the exact same distribution of signals.

Table 6: Main vs Control: Belief Teammate 2 Top by Distibution of Received Signals

	(1) 0 –	(2) 1 –	(3) 2 –	(4) 3 –	(5) 4 –
ATE	−0.016 (0.101)	0.075 (0.099)	0.031 (0.077)	−0.004 (0.096)	0.139** (0.063)
Observations	68	73	91	51	87

Analysis uses nearest neighbor matching with 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. Each column restricts sample to specific distribution of negative signals received (out of 4 total signals).

7 Discussion

The results of this paper point to a high degree of overconfidence, and quite striking patterns of biased, self-serving information processing, when one's own performance is under evaluation. While our focus was on distinguishing precise patterns in attribution, we also found strong positive asymmetry, which fits into a decidedly mixed literature on updating beliefs about ego-relevant events with one-dimensional uncertainty, as discussed by Benjamin (2019). We first briefly discuss potential explanations for why we find strong asymmetry, where others have not.

There are four papers which examine belief updating about relative performance using a similar empirical framework, though applied to updating about one-dimensional uncertainty.[44] Within these papers there are differences on the extent the test is framed as a measure of intelligence or IQ, as well as differences in peer reference groups. This is important, as we should clearly expect motivated cognition biases to be activated more strongly when individuals face a task where their performance sends a strong signal about their innate ability. Presumably, this occurs when a test is framed as an IQ or intelligence test, and when the peer comparison group is larger, thus sending a stronger signal.

In Coutts (2018) and Schwardmann and Van der Weele (2018) the task is not framed as an intelligence test, while in Möbius et al. (2014) and in one of three tasks in Buser et al. (2018) it is. In Coutts (2018) and Möbius et al. (2014) individuals are compared to a large peer reference group, though in the former subjects are asked about beliefs of being in the top 15% which may not be as ego-relevant as the top 50%. Schwardmann and Van der Weele (2018) and Buser et al. (2018) examine smaller reference groups, being in the top 50% of groups of size 4 or 8 respectively.

Coutts (2018) finds negative asymmetry, but is difficult to compare due to lower baseline priors which may activate different cognitive biases with updating than the other papers. Among the remaining three, Schwardmann and Van der Weele (2018) find no evidence of asymmetry, Buser et al. (2018) find evidence of positive asymmetry only when including updates in the wrong direction, and Möbius et al. (2014) find evidence of positive asymmetry in their main results.

As our paper also framed the test as an IQ test, and with a comparison group of 20, this would form a relatively strong signal for subjects. We thus conjecture that ego-relevance may be important in order to activate biases in information processing which may lead to ego-defensive patterns. Future work will be needed to identify whether these conjectures can explain the patterns which have been observed in the literature.[45]

8 Conclusion

How does overconfidence persist in the face of feedback? Psychologists have proposed and tested theories of self-attribution bias, which posit that individuals will be more likely

[44] Möbius et al. (2014), Buser et al. (2018), Schwardmann and Van der Weele (2018), and Coutts (2018).

[45] Beyond this, the context of our experiment was framed in a more natural way, which may also operate in the direction of activation of self-serving biases. We also note that these other studies have not found strong gender differences, like those we find here, including Möbius et al. (2014). However our results are similar to a number of studies which find gender differences in overconfidence, such as Niederle and Vesterlund (2007).

to attribute positive feedback to internal qualities about themselves, and negative feedback to salient external factors. A CEO who faced profit losses does not give a speech about how her company faced "bad luck" or a "random shock", but describes poor economic fundamentals, or blames outspending by her competition.

We took this theory and quantified it, placing it within a simple Bayesian updating framework where individuals face two dimensions of uncertainty. We examined this theory in the context of a natural experiment with two person teams, where we could formally test whether individuals attributed feedback in a biased way between themselves, their teammate, and noise. Our design features two key components. The first is that we utilize a high powered control group, which participates in an identical experiment, but instead faces a team of two random strangers. As such, the control subject's own ability is not relevant for decision making. The second is that we use an indirect belief elicitation procedure that is both incentive compatible, and intuitive.

In our results we document significant evidence of overconfidence. Individuals on average believe they are in the top 50% with a 66% probability. Overconfidence subsequently leads to biased decision making: individuals assign more weight to their own performance relative to both the Bayesian prediction and to those individuals we observe in our ego-irrelevant control. Beyond this we find substantial biases in belief updating when one is a member of the team, our Main treatment. Individuals attribute positive feedback more to themselves at the expense of noise, and to a lesser extent at the expense of their teammate. However the most dramatic bias comes in response to negative feedback, where subjects over-attribute this feedback to noise, and under-attribute it to themselves or their teammate. Notably, in our Control treatment, individuals update symmetrically when receiving positive or negative feedback.

We formalized two types of attribution bias, noisy (NAB), where individuals update neutrally for their teammate but are biased for themselves, and fundamental (FAB), where they update neutrally with respect to noise, but are biased for themselves and their teammate. While neither model fits perfectly, we are able to draw clear conclusions. There is only suggestive evidence of FAB in response to positive feedback, individuals attribute more to their own performance than to their teammate's performance. With negative feedback, FAB is clearly rejected: individuals do not blame others, such feedback is discounted as noise. While we also reject our strict formulation of NAB, subjects do process information in a way which over-attributes negative feedback to noise - they happen to also update in a similarly biased way about their teammate. In this way subjects end up holding more optimistic beliefs both about themselves *and* about their teammates.

This result is confirmed through additional non-parametric tests. Our estimates sug-

gest that after matching individuals in our Main and Control groups on the value of the prior and the sequence of signals observed, those who are updating beliefs about their own performance end up significantly more confident about their ability than those updating about another person. This effect is strongest for those receiving all negative signals. A similar effect of subjects receiving all negative signals is that they end up more optimistic about *their teammate's* performance as well.

Our results provide new insights on belief updating with two-dimensional uncertainty. We document strong evidence of self-attribution biases, and show that mis-attributions can lead individuals to not only be biased about own performance, but also that these biases can spillover to their beliefs about the performance of others. This has important implications, namely that individuals will be less willing to change environments (e.g. teammates), since they will hold inflated expectations about their current environment. It suggests that the self-defeating learning described in Heidhues et al. (2018) may be mitigated since subjects can end up less-pessimistic about the underlying fundamental. Understanding better these implications should be a focus of future research.

References

Arkin, Robert, Harris Cooper, and Thomas Kolditz, "A statistical review of the literature concerning the self-serving attribution bias in interpersonal influence situations," *Journal of Personality*, 12 1980, *48* (4), 435–448.

Azrieli, Yaron, Christopher P Chambers, and Paul J Healy, "Incentives in Experiments: A Theoretical Analysis," *Journal of Political Economy*, 3 2018.

Barber, B M and T Odean, "Boys will be boys: Gender, overconfidence, and common stock investment," *Quarterly Journal of Economics*, 2001, *116* (1), 261–292.

Barron, Kai, "Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains?," *WZB Discussion Paper*, 2017, (October).

Becker, Gordon M., Morris H. Degroot, and Jacob Marschak, "Measuring utility by a single-response sequential method," *Behavioral Science*, 1964, *9* (3), 226–232.

Benabou, R. and J. Tirole, "Self-Confidence and Personal Motivation," *The Quarterly Journal of Economics*, 8 2002, *117* (3), 871–915.

Benabou, Roland and Jean Tirole, "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics*, 2006, *121* (2), 699–746.

Benjamin, Daniel J., *Errors in probabilistic reasoning and judgment biases*, Vol. 2, Elsevier B.V., 2019.

Benoît, Jean Pierre, Juan Dubra, and Don A. Moore, "Does the better-than-average effect show that people are overconfident?: Two experiments," *Journal of the European Economic Association*, 2015, *13* (2), 293–329.

Billett, Matthew T. and Yiming Qian, "Are Overconfident CEOs Born or Made? Evidence of Self-Attribution Bias from Frequent Acquirers," *Management Science*, 2008, *54* (6), 1037–1051.

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

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

Buser, Thomas, Leonie Gerhards, and Jol van der Weele, "Responsiveness to feedback as a personal trait," *Journal of Risk and Uncertainty*, 4 2018, *56* (2), 165–192.

Coutts, Alexander, "Good news and bad news are still news: experimental evidence on belief updating," *Experimental Economics*, 4 2018, pp. 1–27.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, "Investor Psychology and Security Market Under- and Overreactions," *The Journal of Finance*, 12 1998, *53* (6), 1839–1885.

Daniel, Te, Da Seale, and a Rapoport, "Strategic Play and Adaptive Learning in the Sealed-Bid Bargaining Mechanism.," *Journal of mathematical psychology*, 6 1998, *42* (2/3), 133–66.

Deimen, Inga and Julia Wirtz, "A Bandit model of two-dimensional uncertainty," *Working Paper*, 2016.

Doukas, John A. and Dimitris Petmezas, "Acquisitions, Overconfident Managers and Self-attribution Bias," *European Financial Management*, 6 2007, *13* (3), 531–577.

Dubra, Juan and Jean-Pierre Benoît, "Apparent Overconfidence," *Econometrica*, 2011, *79* (5), 1591–1625.

Eberlein, Marion, Sandra Ludwig, and Julia Nafziger, "The Effects of Feedback on Self-Assessment," *Bulletin of Economic Research*, 4 2011, *63* (2), 177–199.

Eil, David and Justin M Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 5 2011, *3* (2), 114–138.

Eliaz, Kfir and Ran Spiegler, "Can anticipatory feelings explain anomalous choices of information sources?," *Games and Economic Behavior*, 7 2006, *56* (1), 87–104.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh, "By chance or by choice? Biased attribution of others outcomes," *Working Paper*, 2019.

Ertac, Seda, "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback," *Journal of Economic Behavior & Organization*, 12 2011, *80* (3), 532–545.

_ **and Balazs Szentes**, "The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence," *mimeo*, 2011.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 2 2007, *10* (2), 171–178.

Fischhoff, Baruch, "Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty.," *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *1* (3), 288–299.

Gervais, Simon and Terrance Odean, "Learning to Be Overconfident," *Review of Financial Studies*, 1 2001, *14* (1), 1–27.

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, 11 1980, *95* (3), 537.

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

Grossman, Zachary and David Owens, "An unlucky feeling: Overconfidence and noisy feedback," *Journal of Economic Behavior & Organization*, 11 2012, *84* (2), 510–524.

Hastorf, Albert H., David J. Schneider, and Judith Polefka, *Person perception*, Reading, Massachusetts: Addison-Wesley Publishing Company, 1970.

Heider, F., "Social perception and phenomenal causality," *Psychological Review*, 1944, *51* (6), 358–374.

Heider, Frtiz, *The psychology of interpersonal relations*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1958.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic Expectations and Misguided Learning," *Econometrica*, 2018, *86* (4), 1159–1214.

Hestermann, Nina and Yves Le Yaouanq, "It's not my fault! Self-confidence and experimentation," 2018.

Hilary, Gilles and Lior Menzly, "Does Past Success Lead Analysts to Become Overconfident?," *Management Science*, 4 2006, *52* (4), 489–500.

Hoffmann, Arvid O.I. and Thomas Post, "Self-attribution bias in consumer financial decision-making: How investment returns affect individuals' belief in skill," *Journal of Behavioral and Experimental Economics*, 2014, *52*, 23–28.

Holt, Charles and Angela M. Smith, "An update on Bayesian updating," *Journal of Economic Behavior & Organization*, 2 2009, *69* (2), 125–134.

Hossain, Tanjim and Ryo Okui, "The Binarized Scoring Rule," *The Review of Economic Studies*, 1 2013, *80* (3), 984–1001.

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

Kelley, Harold H., "The processes of causal attribution.," *American Psychologist*, 1973, *28* (2), 107–128.

Kim, Y. Han (Andy), "Self attribution bias of the CEO: Evidence from CEO interviews on CNBC," *Journal of Banking and Finance*, 2013, *37* (7), 2472–2489.

Köszegi, B and M. Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 11 2006, *121* (4), 1133–1165.

Larrick, Richard P., Katherine A. Burson, and Jack B. Soll, "Social comparison and confidence: When thinking youre better than average predicts overconfidence (and when it does not)," *Organizational Behavior and Human Decision Processes*, 1 2007, *102* (1), 76–94.

Lassiter, G Daniel, Andrew L Geers, Patrick J Munhall, Robert J Ploutz-snyder, and David L Breitenbecher, "Illusory Causation: Why It Occurs," *Psychological science*, 2002, *13* (4), 299–306.

Li, Feng, "Managers Self-Serving Attribution Bias and Corporate Financial Policies," *SSRN Electronic Journal*, 2010.

Libby, Robert and Kristina Rennekamp, "Self-Serving Attribution Bias, Overconfidence, and the Issuance of Management Forecasts," *Journal of Accounting Research*, 2011, *50* (1), 197–231.

Machina, Mark J, ""Expected Utility" Analysis without the Independence Axiom," *Econometrica*, 1982, *50* (2), 277–323.

Malmendier, Ulrike and Geoffrey Tate, "Does Overconfidence Affect Corporate Investment? CEO Overconfidence Measures Revisited," *European Financial Management*, 11 2005, *11* (5), 649–659.

Marín, C and A Rosa-García, "Gender bias in risk aversion: evidence from multiple choice exams," 2011.

Mezulis, Amy H., Lyn Y. Abramson, Janet S. Hyde, and Benjamin L. Hankin, "Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias.," *Psychological Bulletin*, 2004, *130* (5), 711–747.

Miller, Dale T. and Michael Ross, "Self-serving biases in the attribution of causality: Fact or fiction?," *Psychological Bulletin*, 1975, *82* (2), 213–225.

Möbius, M M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing Self-Confidence," *mimeo*, 2014, pp. 1–43.

Moore, D.A. and P.J. Healy, "The Trouble With Overconfidence," *Psychological Review*, 2008, *115* (2).

Moore, Don A. and Deborah A. Small, "Error and bias in comparative judgment: On being both better and worse than we think we are.," *Journal of Personality and Social Psychology*, 2007, *92* (6), 972–989.

Niederle, M. and L. Vesterlund, "Do Women Shy Away From Competition? Do Men Compete Too Much?," *The Quarterly Journal of Economics*, 8 2007, *122* (3), 1067–1101.

Pekrun, Reinhard and Herbert W. Marsh, "Weiners attribution theory: Indispensable-but is it immune to crisis?," *Motivation Science*, 2018, *4* (1), 19–20.

Pryor, John B. and Mitchel Kriss, "The cognitive dynamics of salience in the attribution process," *Journal of Personality and Social Psychology*, 1977, *35* (1), 49–55.

Pulford, Briony D. and Andrew M. Colman, "Overconfidence: Feedback and item difficulty effects," *Personality and Individual Differences*, 7 1997, *23* (1), 125–133.

Schwardmann, Peter and Joel Van der Weele, "Deception and Self-Deception," 2018.

Shaver, Kelly, *An Introduction to Attribution Processes*, London: Routledge, 1983.

Silvia, Paul J. and T. Shelley Duval, "Predicting the Interpersonal Targets of Self-Serving Attributions," *Journal of Experimental Social Psychology*, 2001, *37* (4), 333–340.

Spilka, Bernard, Phillip Shaver, and Lee A. Kirkpatrick, "A General Attribution Theory for the Psychology of Religion," *Journal for the Scientific Study of Religion*, 1985, *24* (1), 1.

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

Tetlock, Philip E. and Ariel Levi, "Attribution bias: On the inconclusiveness of the cognition-motivation debate," *Journal of Experimental Social Psychology*, 1982, *18* (1), 68–88.

Tversky, Amos and Daniel Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, 9 1973, *5* (2), 207–232.

Weiner, Bernard, "Attribution Theory," in "A Companion to the Philosophy of Action," Oxford, UK: Wiley-Blackwell, 7 2010, pp. 366–373.

_ **and Sandra Graham**, "Attribution in personality psychology.," in "Handbook of personality: Theory and research, 2nd ed.," New York, NY, US: Guilford Press, 1999, pp. 605–628.

Wozniak, David, William T Harbaugh, and Ulrich Mayr, "The effects of free and costly feedback on gender differences in competitive choices . The effects of free and costly feedback on gender differences in competitive choices.," *mimeo*, 2016.

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *Presentation for ECBE*, 2017.

Zuckerman, Miron, "Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory," *Journal of Personality*, 6 1979, *47* (2), 245–287.

A Chosen Weights

While the primary focus of the empirical analysis is on determinants of beliefs and belief updating, it is informative to investigate how beliefs and updating affect subject's weighting decisions. Recall that individuals had to choose a weight from 0 to 1, with 0 representing all of the weight on teammate 2, and 1 representing all of the weight on teammate 1. Recall the theoretical prediction is that the weight chosen should be invariant to feedback. That is, after controlling for the initial weight, neither positive nor negative feedback should alter the submitted weight.

Table A.1 shows regressions which examine impacts of subject characteristics and the main treatment on weighting decisions. The theoretical prediction is that the initial weight should have a coefficient of one, and all other coefficients should be zero. From the table one can see that this is not the case. While the initial weight is positive and significant, it is less than one. What is more interesting is that against the theoretical predictions, positive feedback has a statistically significant effect on the weight chosen, in columns (1) and (2). Additionally, there is some evidence that being a member of the team has a statistically significant effect on the chosen weight.

Yet, as columns (3) and (4) show, the positive effect of both a positive signal and the treatment are coming from the interaction between the two. In particular, this interaction increases the weight by 6.4 percentage points. This is about an 11% increase on the average weight chosen. Thus, when individuals are part of the team, when receiving a positive signal they increase the weight on their own performance by 6.4 percentage points, despite the theoretical benchmark being to not alter the weight. This result is consistent with the results on updating. In the Control treatment the responsiveness to feedback was balanced across both teammate 1 and 2, and for both positive and negative feedback. In contrast, in the Main treatment, positive feedback was attributed more to teammate 1, while negative feedback was not attributed to either. As such, one would expect the result of Table A.1 which shows that positive signals in the Main treatment result in a higher weight on teammate 1.

Table A.1: Submitted Weight on Teammate 1

	(1)	(2)	(3)	(4)
Initial Weight	0.600^{***}	0.515^{***}	0.518^{***}	0.473^{***}
	(0.033)	(0.042)	(0.042)	(0.044)
+ Signal	5.435^{***}	5.364^{***}	2.367	0.521
	(1.458)	(1.435)	(2.136)	(2.081)
Main Treatment	3.540	4.267^{*}	1.210	0.854
	(2.320)	(2.273)	(2.843)	(2.726)
+ Signal × Main Treatment			5.982^{**}	6.435^{**}
			(2.833)	(2.738)
Female	2.767	2.223	2.480	2.244
	(2.271)	(2.194)	(2.179)	(2.143)
Age	-0.387	-0.409^{*}	-0.410^{*}	-0.381
	(0.237)	(0.241)	(0.241)	(0.236)
# Attempted by Teammate 1		2.976^{***}	2.965^{***}	1.675^{**}
		(0.626)	(0.626)	(0.687)
# Attempted by Teammate 2		-1.272^{**}	-1.276^{**}	-1.675^{***}
		(0.558)	(0.555)	(0.528)
Score of Teammate 1 on IQ Test				0.608^{***}
				(0.158)
Round Fixed Effects	✓	✓	✓	✓
R^2	0.38	0.40	0.40	0.42
Observations	2595	2595	2595	2595

Analysis uses OLS regression. Difference is significant from 0 at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level.

Finally we briefly examine why some individuals choose to submit different weights from those recommended; recall that this was 11% of observations. Among those who submitted a different weight, the average difference from the optimal was 0.056 (recalling that $\omega \in [0, 1]$). However there are no differences by treatment. Thus we do not see any significant patterns here. However, of note is that among individuals who elected to choose a non-recommended weight, there are significant differences in the beliefs they reported about teammate 2. In particular, those who chose higher than recommended weights reported on average greater beliefs about teammate 2 being in the top half, by a significant 22 percentage points (Ranksum p-value 0.000). There were no differences by beliefs of teammate 1. In other words, individuals who saw a lower weight on teammate 1's performance, were more likely to intervene and select a higher weight. However this pattern is not driven by own performance, as it is equally present in Main and Control.

B WTP

Recall that in Wave 2 we provided subjects with the opportunity to be randomly rematched to a new teammate 2, using the BDM mechanism. Subjects i could bid $x_i \in$ €$[0, 5]$, where €5 is the risk-neutral maximum value of switching.[46] After submitting their bid, the computer randomly generated a price, $p \in [0, 1]$ using a continuous distribution.[47] Whenever $x_i > p$ they would pay the price p out of their earnings, and be matched with a new teammate. If $x_i \leq p$ they would not pay anything, and stay matched with the same teammate.

Given the reported beliefs of subjects we are able to calculate whether it would be optimal for them to change teammates, assuming risk neutrality. For the 231 subjects in Wave 2, after four rounds of feedback, 70 subjects had a positive value of changing teammates. The average expected value from changing to a new teammate 2 was €1.42, ranging from €0 to €5.

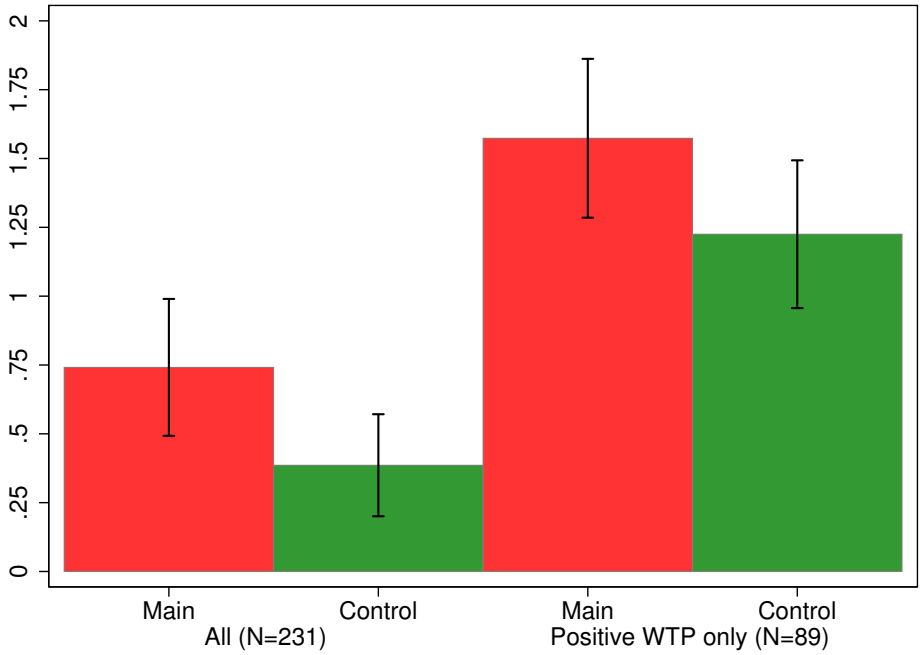
In fact, 89 subjects expressed a positive WTP. Of the 70 subjects with positive expected value, 40 of them actually stated a positive WTP, on average equal to €1.64, not far from the true expected value of €1.42. The remaining 49 subjects expressed a positive WTP despite having either 0 value for switching (15) or negative value (34).

Regarding whether there are differences across Main and Control, 40 of the 89 were in Main (predicted 33) and 49 were in Control (predicted 37). Thus there are no striking patterns. However, we note that there are differences in the overall WTP across Main and Control, driven by those in Main submitting lower WTP. These can be seen in Figure B.1. Average WTP in Main is €0.39, while in Control it is €0.74, significantly different at the 1% level (Ranksum p-value 0.006). Restricting the sample only to positive WTP, the Ranksum p-value is 0.132, $N = 89$. However, some of this gap is accounted for by actual belief differences (in particular the calculated optimal WTP is slightly lower in Main, given beliefs). Thus while there is suggestive evidence of lower WTP in Main treatment relative to Control, even after accounting for beliefs, we do not have the power to show this statistically.

[46] Note that the worst outcome for subjects is when both teammates are in the bottom half, where they will earn €0 with certainty. If one is in the top half, they can select ω accordingly to ensure a high probability of earning €10. Since there is a 50% probability a randomly selected person is in the top half, the expected value of being matched with them is €5.

[47] In fact we used a left skewed distribution by combining two uniform distributions, in order to generate more re-matching.

Figure B.1: Willingness to pay



WTP in Euros of subjects to change teammate 2. Left side includes all data, right side includes only positive values of WTP. Wave 2 only. 95% confidence intervals shown.

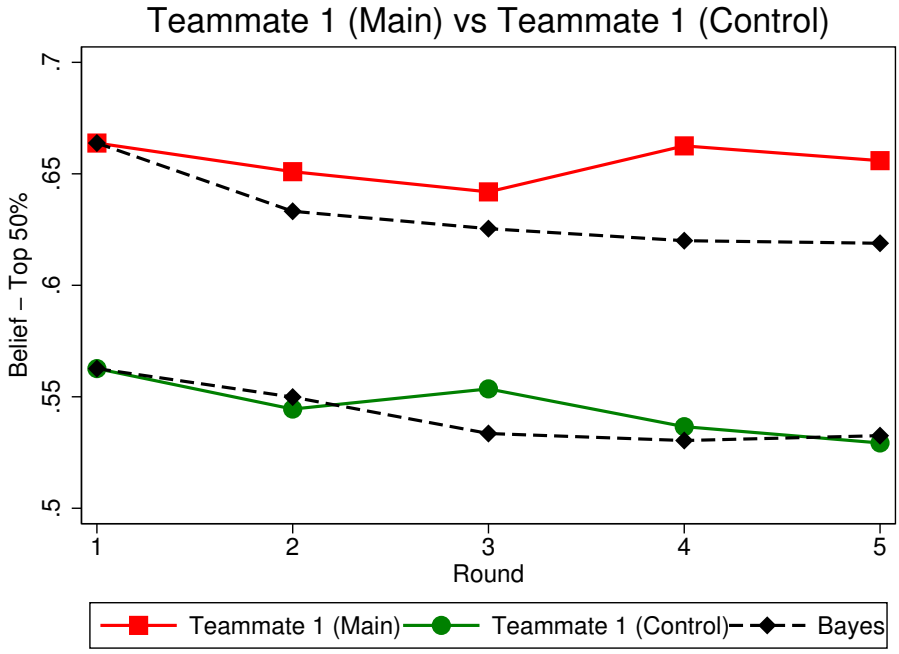
C Examining Posterior Beliefs

Figures C.1 and C.2 examine the evolution of beliefs in response to feedback for teammate 1 and 2 respectively, starting from the first prior, before receiving any feedback. While posterior beliefs about one's self (Main, teammate 1) are significantly greater than beliefs about teammate 1 in the Control, this is in large part driven by differences in prior beliefs due to overconfidence. In both Figures one can see a pattern that posterior beliefs in the final round deviate further from the Bayesian prediction in Main compared to Control, both for teammate 1 and 2.

Figure C.3 examines this more closely, presenting the difference between reported posteriors and the Bayesian prediction given subjects' initial priors, after four rounds of feedback. This corresponds to round 5 in the two figures above. While this does present evidence that positive deviations are more pronounced in the Main treatment, we also note that the difference between the deviations in Main and Control are not significantly different at conventional levels.

Figures C.4-C.7 present Epanechnikov kernal-weighted local polynomial smoothing plots regarding the relationship between priors and posteriors. The sample is identical

Figure C.1: Evolution of Beliefs: Teammate 1



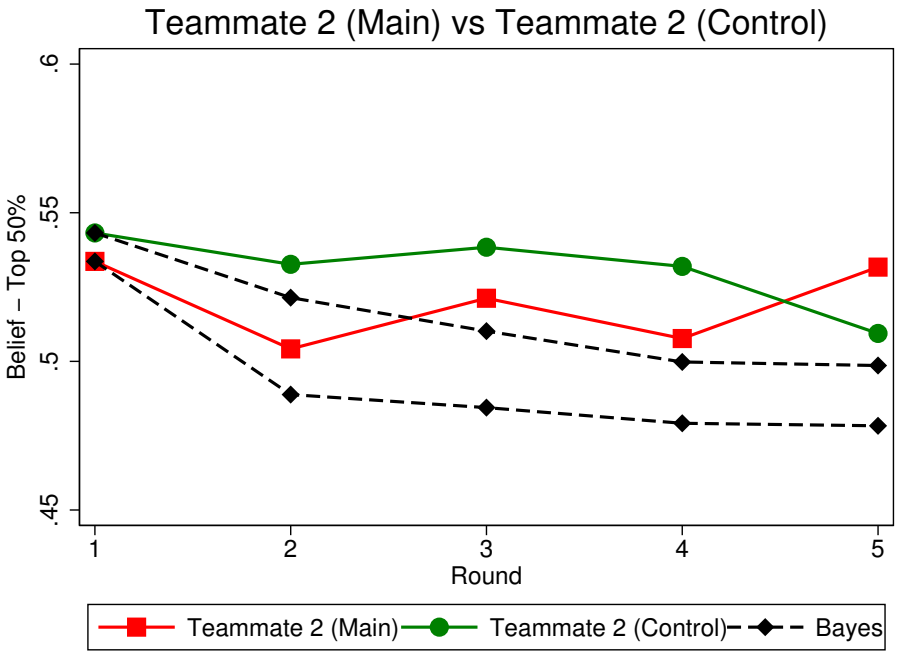
Evolution of beliefs about teammate 1 starting from prior beliefs with 4 round of feedback. Bayesian benchmark is calculated from subject's first prior, then evolves given actual signals observed. Standard error bars omitted for clarify (error bars are always overlapping with bayesian predictions).

to that of Tables 1 and 2. The Bayesian estimates are presented in black, while Main treatment estimates are red, and Control are green. Shaded 95% confidence intervals are shown for each case.

For teammate 1, after receiving positive feedback, updating is similar, though there are significant differences for larger priors, in that individuals updating about their own ability update more in response to positive signals, relative to the control treatment. After receiving negative feedback, a similar pattern emerges; individuals update less in the negative direction in the Main compared to Control. This occurs predominately for priors less than 50%.

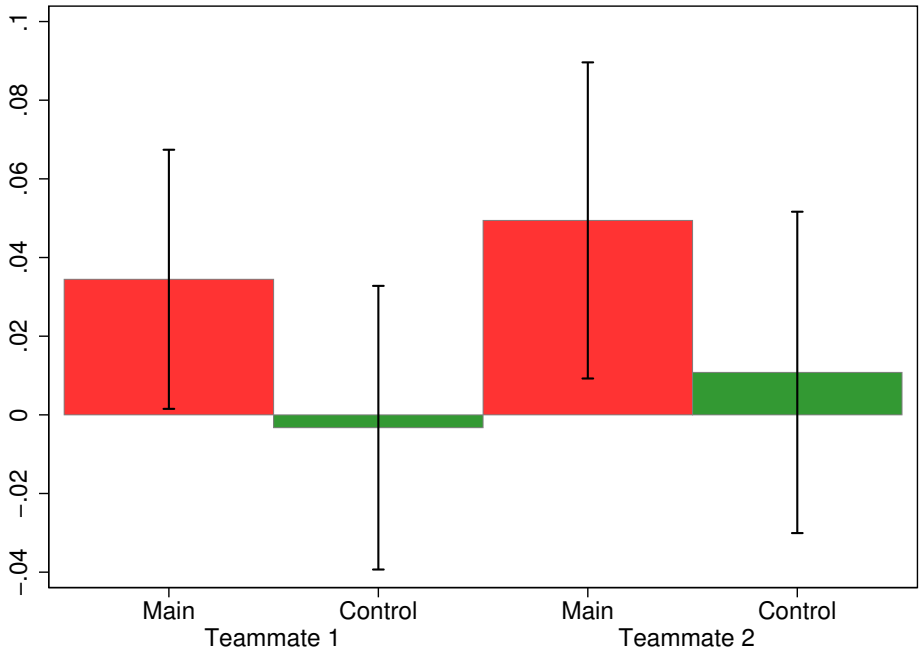
Finally regarding teammate 2, the patterns similar but not as pronounced. After a positive signal there are not clear differences between Control versus Main. For a negative signal there it appears that subjects in Main update less in response to a negative signal compared to those in Control, with these patterns present only for priors less than 50%, as with teammate 1.

Figure C.2: Evolution of Beliefs: Teammate 2



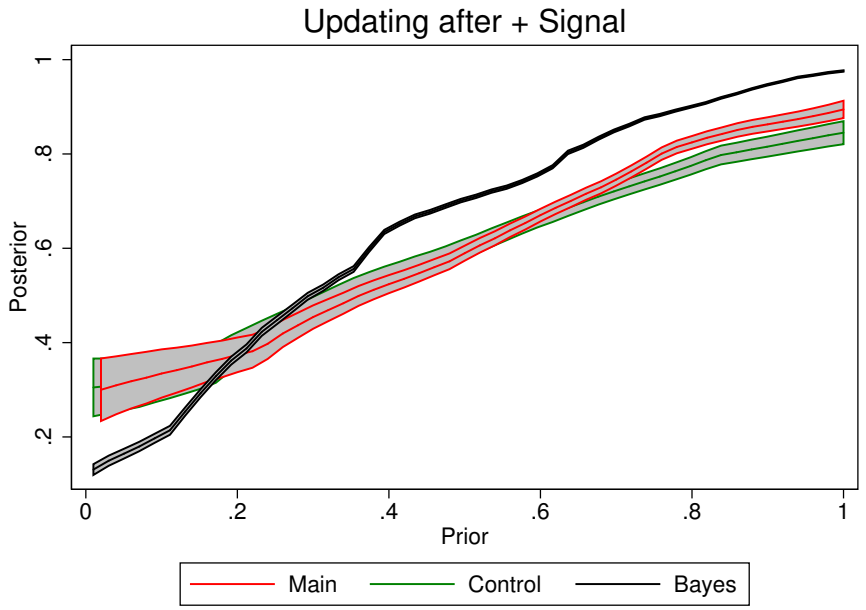
Evolution of beliefs about teammate 2 starting from prior beliefs with 4 round of feedback. Bayesian benchmark is calculated from subject's first prior, then evolves given actual signals observed. Standard error bars omitted for clarify (error bars are always overlapping with bayesian predictions).

Figure C.3: Raw Deviation of Posterior Beliefs from Bayesian Benchmark



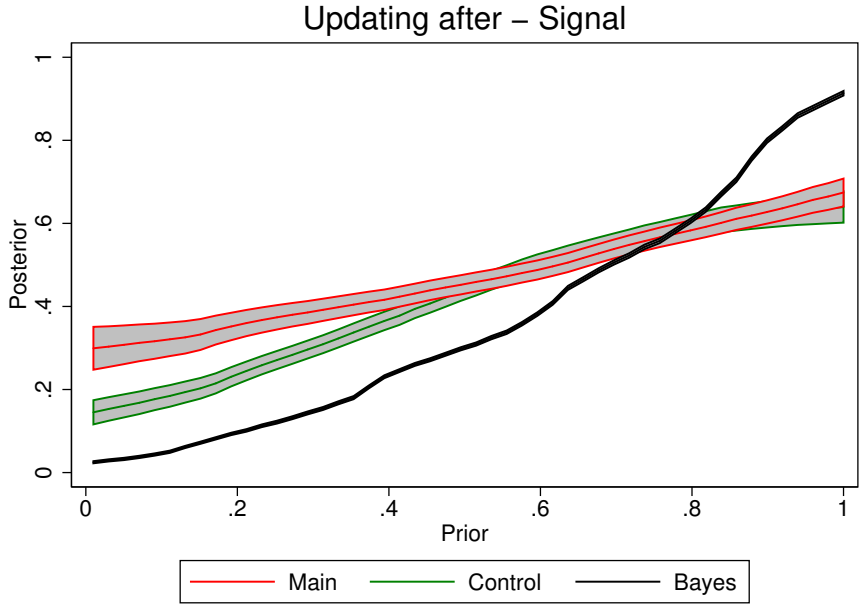
Plot of the difference between Posterior beliefs and Bayesian beliefs after 4 rounds of feedback. Bayesian beliefs are calculated using subject priors before any feedback.

Figure C.4: Teammate 1: Positive Feedback



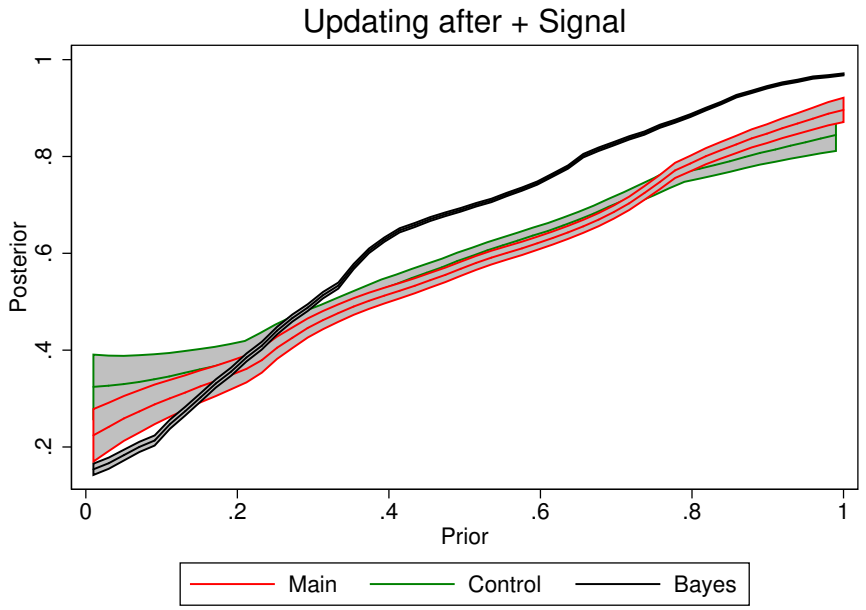
Epanechnikov kernal-weighted local polynomial smoothing plot showing relationship between priors and posteriors for teammate 1 in response to positive feedback. Sample includes Parts 2 and 3, with the same sampling restrictions as Table 1.

Figure C.5: Teammate 1: Negative Feedback



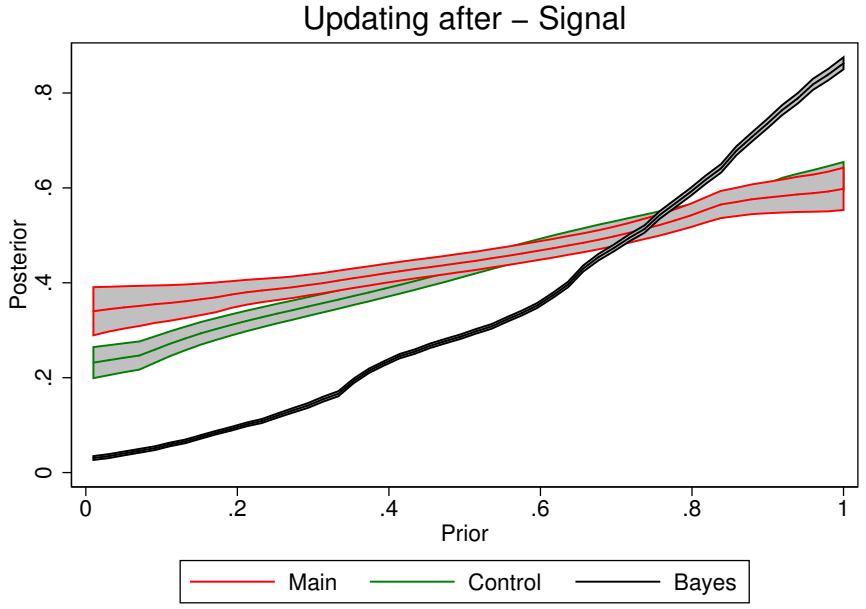
Epanechnikov kernal-weighted local polynomial smoothing plot showing relationship between priors and posteriors for teammate 1 in response to negative feedback. Sample includes Parts 2 and 3, with the same sampling restrictions as Table 1.

Figure C.6: Teammate 2: Positive Feedback



Epanechnikov kernal-weighted local polynomial smoothing plot showing relationship between priors and posteriors for teammate 1 in response to positive feedback. Sample includes Parts 2 and 3, with the same sampling restrictions as Table 2.

Figure C.7: Teammate 2: Negative Feedback



Epanechnikov kernal-weighted local polynomial smoothing plot showing relationship between priors and posteriors for teammate 1 in response to negative feedback. Sample includes Parts 2 and 3, with the same sampling restrictions as Table 2.