No one to blame: Self-attribution bias in updating with two-dimensional uncertainty*†

Alexander Coutts Leonie Gerhards Zahra Murad

June 28 2019

Abstract

We investigate how overconfidence persists in the face of objective feedback which depends on two dimensions of uncertainty. Self-attribution biases exist when individuals take credit for good outcomes, but blame poor outcomes on external factors. 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 which updates for a third party. Moreover, self-serving biases spill over to positively affect beliefs about the teammate, suggesting that nurturing such biased beliefs can generate a more broadly distorted worldview.

*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, Bayesian Crowd Conference, briq Workshop on Beliefs, CEA Banff, IMEBESS Utrecht, ESA Berlin, HEC Lausanne, Lisbon Game Theory Meetings, M-BEES, NYU CESS, NYU Shanghai, University of Portsmouth, SHUFE, THEEM, and WZB. We gratefully acknowledge financial support from the Hamburgische Wissenschaftliche Stiftung and the University of Hamburg.

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. 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. In this paper we focus on how overconfidence may persist through biased information processing about self-relevant information, specifically when this information comes bundled with an additional source of uncertainty.[1]

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 (Mezulis et al., 2004). 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]

The study of self-attribution bias in social psychology has focused on how different features of the environment interact to enable attributions which lead to biased self-assessments. Attribution may occur vis-a-vis some external fundamental, e.g. the state of the economy or one's colleagues, or it may occur through attribution towards more unstable/idiosyncratic factors such as chance, e.g. the student above blaming bad luck for the group's low grade. In this paper we present two modified Bayesian theories of self-attribution bias, and test them with a corresponding experiment. In doing so we operationalize a concept that has been the subject of many psychological studies, but which has typically been studied without a quantitative, falsifiable, model.[3] Among the theories, one is focused on mis-attribution towards external fundamental factors, deemed fundamental attribution bias (FAB) while the other is focused on mis-attribution towards idiosyncratic noise, deemed noisy attribution bias (NAB).

Distinguishing the two models has broad implications for decision making: for both biases, individuals end up with overly positive views about personal qualities, but with FAB they additionally end up with overly negative assessments of an external fundamental. FAB thus leads to a double

[1] We abstract away from 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 Bénabou and Tirole (2002), could lead to overconfidence if individuals recall information in a biased way; Zimmermann (2019) in fact finds 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 Bénabou and Tirole (2002) and our later discussion.

[3] This reflects differences in methodological approaches, as these theories regarding attribution biases are typically non-quantitative. Despite their differences, these studies in social psychology have been invaluable in identifying and contributing to our understanding of these biases, and they form the motivation for this paper.

penalty for decision-making: individuals are biased not only by rosy perceptions of their self, but of gloomy perceptions about other features of the environment. A student suffering from FAB may both (sub-optimally) abandon her old group, and may take biased overconfident decisions within a new group.[4]

Beyond the importance of these implications, we put particular emphasis on studying and measuring FAB for three additional reasons. First, the bias is referenced in numerous studies within economics and finance, e.g. Daniel et al. (1998a) and Gervais and Odean (2001), but corresponding direct quantitative evidence is scarce.[5] Second, work in social psychology has emphasized features of attribution bias which suggest a greater role for FAB than NAB. Early work by Heider (1958) highlighted the enhanced attention paid to more stable factors in causal attribution. Later empirical work corroborated this intuition, but highlighted the role of salience or availability bias (Tversky and Kahneman, 1973) in generating attribution to more stable fundamental factors, see Lassiter et al. (2002). The third reason derives from an important contribution of our theory, which shows that FAB can uniquely be derived starting from a micro-founded model which trades-off the benefits and costs of biased information processing.

Our lab experiment tests these two theories of self-attribution bias using a context whereby a two-person team's output depends on the ability of both members, measured through an IQ-style test. 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; however the feedback from these two sources cannot be disentangled. Payoffs in the experiment are such that individuals are incentivized in a novel, but intuitive way, to truthfully report their beliefs about the probabilities that each of the two teammates scored in the top half on the test compared to a reference group. Given these beliefs, the computer then optimally selects how much to weight one teammate's performance relative to the other teammate's performance, by maximizing the expected probability of earning €10 in the experiment.

Our experiment is the first to empirically study updating with two dimensions of uncertainty using a structural framework. Beyond this we utilize a fully powered control group, which participates in an identical experiment but without own test performance being relevant. Subjects in this control group play the role of a randomly chosen "teammate 1" (matched with a differently randomly chosen teammate 2), and observe feedback about this team, that is unrelated to their own ability. Throughout, we refer to the experimental sessions involving subjects themselves being teammate 1 as the Main treatment, while we refer to our control group making third party decisions for a different teammate 1 as the Control treatment.

Our results are as follows. In the Main treatment, individuals update in a significantly biased way about their own ability, incorporating positive feedback more than negative feedback, i.e. positive

[4] An extreme example of how learning could be self-defeating 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. On the other hand, if individuals routinely have opportunities to change their environments, in the long run FAB may create more opportunities for biased individuals to learn, as in Hestermann and Le Yaouanq (2018).

[5] Other empirical studies that reference FAB include: Billett and Qian (2008), Doukas and Petmezas (2007), Hilary and Menzly (2006), Hoffmann and Post (2014), Kim (2013), Li (2010), and Libby and Rennekamp (2011).

asymmetry. By contrast, in the Control treatment, subjects incorporate feedback about teammate 1 in a symmetric way. This is consistent with the theories of both FAB and NAB. However, in contrast to either of these theories, individuals also exhibit positive asymmetry in updating beliefs about their teammate in the Main treatment. While this asymmetry is not as strong as when updating about own performance, it is absent in the Control treatment.

As FAB predicts negative asymmetry for updating about the teammate (mis-attribution across self and teammate), while NAB predicts no such asymmetry (mis-attribution across self and noise), these results are not consistent with either model. Yet positively biased beliefs about both teammates lead to more moderate weighting choices, which mitigate material losses from overconfidence in the experiment. To better understand whether such behavior is optimal, we explore variations of a micro-founded model of FAB. Allowing for the possibility that individuals can process information in a biased way about both teammates, we find that indeed positive asymmetry for updating about both teammates may be optimal. By lowering the material costs to overconfidence, these patterns of updating can enable even higher self-serving beliefs.

Finally, while our experimental tests of the theory require that we impose a quasi-Bayesian structure on the analysis, the existence of our Control experiment also allow us to make non-parametric comparisons. Using a matching strategy we match our Main and Control experiments on prior beliefs. We show that posteriors in the Main group (when updating is about own performance) are 8.4 percentage points higher after 4 rounds of feedback, conditional on having the same initial priors and having received the exact same feedback. These differences are strongest for individuals receiving mostly negative signals, consistent with our structural analysis.

Overall we present robust evidence that individuals update differently about their self and their teammate when information is self-relevant. In understanding the mechanisms behind this result, we find that motives for self-enhancement can drive positive asymmetric updating about both their own and their teammate's performances. Our theoretical analysis and results generate new insights in relation to an emerging literature on learning with two dimensions of uncertainty. Specifically, we document how motivated cognition can spill over to distort beliefs about non-ego relevant states of the world. In line with this, we find that individuals in our Main treatment are significantly less likely to opt to switch to a new teammate, due to inflated beliefs about their current teammate.

The rest of the paper proceeds as follows. After a literature review, we outline our experimental context and design. This is followed by our theory, which focuses on the distinctions between NAB vs FAB. We finally describe our predictions, followed by results, and conclude with a short discussion.

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. Dubra and Benoît (2011) noted that rational behavior may generate overconfident-appearing

data. Yet even accounting for this, many studies have found evidence consistent with overconfidence, see Benoît et al. (2015), and the discussion contained therein.

A broad literature within economics has emerged on motivated cognition, which explores the motivations for holding self-serving beliefs, and can potentially account for the presence and persistence of overconfidence. The benefits to overconfidence may arise from (i) direct utility from holding overconfident beliefs (Möbius et al., 2014; Brunnermeier and Parker, 2005) for example arising from self-esteem or ego-protection, (ii) benefits to personal motivation or self-signalling (Bénabou and Tirole, 2002, 2009, 2011), or (iii) strategic signalling motives/persuasion of others (Burks et al., 2013; Schwardmann and Van der Weele, 2018). These three explanations have long been a part of the core motivation for attribution theory of social psychology, corresponding to (i) self-enhancement/protection (ii) positive presentation of self to others, and (iii) belief in effective control; see Kelley and Michela (1980) and Tetlock and Levi (1982).

A long-standing literature from social psychology has suggested that the supply of such overconfident beliefs may come from self-serving biases in the attribution process, i.e. a tendency to "attribute success to our own dispositions and failure to external forces" (Hastorf et al., 1970) (p.73). Such self-attribution bias has its origins in the writings of Fritz Heider. Heider (1944, 1958) described the innate human desire to to explain behaviors and outcomes, further noting 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). Empirical evidence on this can be found in Pryor and Kriss (1977) with further discussion in Lassiter et al. (2002).

The resulting theories of attribution focus more on general principles rather than tractable models, as discussed in Kelley (1973) and Weiner (2010), and many studies of the bias are empirical in nature. While these studies are provoking, concerns arose from this literature on whether the researchers' theories matched how subjects actually perceived the experimental environments they faced.[6] 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).

One of the few papers which bridges these psychological motivations with economic theory is Bénabou and Tirole (2009), who do so in the context of a partnership of two individuals who observe high or low joint output, and benefit from preserving their self image about who was responsible for output. While there are parallels with our team setting, the mechanism they highlight (self-signalling through imperfect recall) differs from our current setting, which focuses on belief distortion directly.[7]

Our analysis of the measurement and evolution of beliefs in a quasi-Bayesian framework presents some challenges for connecting previous theory and evidence within social psychology. The identity

[6] In an archetypal experiment in social psychology studying self-attribution in achievement tasks, individuals are given a task (e.g. anagrams task), then receive success or failure feedback (sometimes falsified), and are asked to distribute responsibility for this outcome among internal vs external factors (Miller and Ross, 1975; Mezulis et al., 2004). Another concern was in interpreting responses to feedback that had been falsified. 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.

[7] In a related paper, Bénabou and Tirole (2011) consider a broader framework for analyzing decision making when beliefs are valued by individuals and can be viewed as investments.

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 quantitative framework of self-attribution bias directly speaks to a growing literature in economics which studies belief updating and motivated cognition, yet we also maintain the salient structures emphasized within psychology. We focus our theoretical discussion and experiment on having a human teammate's ability to serve as a salient source of uncertainty. 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. Re-matching individuals with new teammates every period would reduce stability, and would be predicted to result in less attribution.

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. Our focus on updating with two dimensions of uncertainty connects the literature on one dimensional ego-relevant updating with an emerging theoretical literature on learning and decision making with multiple dimensions of uncertainty. This former strand of literature on one dimensional uncertainty in ego-relevant settings includes Buser et al. (2018), Coutts (2019a), Eil and Rao (2011), Ertac (2011), Grossman and Owens (2012), Möbius et al. (2014), and Schwardmann and Van der Weele (2018). 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.[8] 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 Brunnermeier and Parker (2005). 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).[9] 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 the overlap with each paper in turn, when possible we discuss their theory within 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 envi-

[8] 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, 2019a; 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 Barron (2017), Burks et al. (2013), Eberlein et al. (2011), Erkal et al. (2019), Pulford and Colman (1997), Ertac and Szentes (2011), and Wozniak et al. (2014).

[9] 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.

ronment, 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. 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". 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.[10] They demonstrate that under certain conditions, an overconfident individual will perceive poor outcomes as reflecting poor performance by another teammate rather than herself. In response, they show that the individual decision making can lead to a cycle of self-defeating learning, and poor outcomes which the agent increasingly attributes to her teammate.

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 framework 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 weighting decisions and outcomes.

3 Experimental Design

3.1 Overview

The experiment was conducted at the WiSo experimental laboratory at the University of Hamburg. A total of 426 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.[11]

The experiment consisted of two main parts. Part 1 consisted of a 10 minute IQ style test which allowed subjects to be ranked according to their performance. Part 2 consisted of matching subjects into two-person teams, and eliciting subjects' beliefs about their own and their teammate's relative performance in Part 1. Subjects received four rounds of feedback, and reported their beliefs five times

[10] 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.

[11] 52% of the participants were Female. 192 subjects participated in the first wave, while 234 subjects participated in the second. In one session of wave 2 a fire alarm went off at the end, invalidating only data for Part 3. Due to a small glitch, some subjects inadvertently skipped entering beliefs, which leaves us with 3155 out of 3170 observations. Earnings included a €5 show-up fee.

in total. Part 3, only present in the second wave, asked subjects' willingness to pay to switch their teammates, and involved a further four rounds of feedback with the (potentially new) teammate.

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 was remaining on 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."[12]

Subjects were assigned one of two versions of the test, randomized at the session level, one harder and one easier. This allowed us to examine whether the hard-easy effect was present in our setting.[13] 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. These incentives ensured that subjects attempted a question only if they were relatively sure that they knew the answer. This way we ensured that the attempted number of questions would carry some informational value, which we use in the later parts of the experiment.[14] 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 both, payoffs were determined by subjects' reported beliefs which resulted in a payoff-relevant weighting decision, explained in more detail in Section 3.4. The only difference was that in the Main treatment subjects' own performance was relevant, while in the Control, it was not.

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 2. After the elicitation, subjects continued to Part 3 which was the same as Part 2, but with possibly a new teammate. Below we describe the components of the experiment in more detail. Full experimental instructions are presented in the Online Appendix Section 7 and Table 1 presents the flow of the experiment.

3.2 Main Treatment

At the beginning of Part 2, subjects were paired into teams of two. Pairings remained constant throughout this part. Their individual performances on the test from Part 1 jointly defined their "team

[12] 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 with a stated use of measuring general intelligence. However, the scores achieved cannot be interpreted as true IQ estimates.

[13] See Larrick et al. (2007) and Moore and Small (2007). The hard-easy effect stipulates that individuals will be more upwardly biased in estimates of their relative performance on easy rather than hard tasks. Our interest was also in examining whether there were hard-easy differences in information processing.

[14] 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 evidence for this.

Table 1: Experimental Flow

Part 1	• IQ task (10 minutes) • Piece rate paid: wrong answers penalized
Part 2	• Teammate 1 is matched at random to a teammate 2 • Observe # of attempted questions for teammate 2 • Report prior beliefs about teammate 1 and teammate 2 • Submit first weight **Repeated $\times$ 4 times:** • Receive feedback • Report posterior beliefs about teammate 1 and teammate 2 • Submit the weight
Part 3: **Wave 2 only**	• Willingness to pay to switch teammate 2 • Lottery determines whether teammate 2 is switched or not • Observe # of attempted questions for (new) teammate 2 • Report beliefs about teammate 1 and teammate 2 • Submit the weight **Repeated $\times$ 4 times:** • Receive feedback • Report posterior beliefs about teammate 1 and teammate 2 • Submit the weight

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 test from Part 1. Number of attempted questions provides some limited information about teammate's performance, which generated variation in prior beliefs. This will be useful for our analysis which matches Main and Control 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 teammate belonged to the top or bottom half until the end of the experiment. Not comparing teammates' scores to each other, but to the same comparison group, ensured that the teammates' individual rankings were independent of each other.

3.3 Control Treatment

In the Control treatment, subjects play the role of a third party who must report beliefs 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 submitted 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 had identical information about all decision-relevant variables as the subjects in the Main treatment (who were themselves in the role of teammate 1).

3.4 Weighting Decision and Belief Elicitation

Subjects were informed 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. The weighting decision depended on subjects' reported beliefs and only affected subjects' own earnings which was emphasized in the instructions. This ensured that social preferences played no role in subjects' decisions.

Subjects' payoffs involved the probability of earning a positive payment $P = €10$, and depended on four relevant performance states, the 2×2 set which corresponds to each teammate being either in the top or bottom half. We denote these four states by S_1S_2, where $S_i \in \{T, B\}$ for top or bottom respectively, for teammate i. Subjects would earn an amount of €10 (€0) for sure, if both of the teammates were ranked in the top half TT (bottom half BB) in Part 1. 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 a weighting decision $\omega_t \in [0, 1]$. Specifically, the probability of earning €10 was given by $\sqrt{\omega_t}$ if teammate 1 scored in the top half and teammate 2 in the bottom half (TB) and $\sqrt{1-\omega_t}$ if teammate 1 scored in the bottom half and teammate 2 in the top half (BT). Thus, only these two latter states are payoff relevant for the weighting decision. Based solely on the subjects' reported beliefs, the computer calculated the optimal weight and advised subjects in an intuitive way, using graphical tools and an explanation of which weight would give them the highest probability of winning €10.[15]

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 probabilities of the two teammates scoring in the top half. Subjects were given complete information about this structure of expected payoffs. They were then told that the computer would maximize the probability that they earned the €10 given the beliefs that they reported. After they entered their beliefs, the screen displayed the true (ex-ante) probability of winning the €10 for every possible weight $\omega_t \in [0, 1]$. It

[15] Subjects saw a transformed weight from 0 to 100 rather than 0 to 1, to make the experiment more intuitive.

highlighted the weight that gave them the maximum probability of winning, however they were free to ignore this recommendation and enter any other weight. The optimal weight is independent of risk preferences, see Section 4.2.

The key to understanding the incentive compatibility of our procedure is that the weighting decision affects only the probability of earning a fixed prize of €10. Assuming subjects can form subjective beliefs, as long as they strictly prefer a higher probability of earning €10, it is in their best interest to truthfully report those beliefs. This procedure is thus novel in its indirect implementation, but shares the same incentive compatibility properties of other elicitation procedures such as matching probabilities (Holt and Smith, 2009; Karni, 2009), or the binarized scoring rule (Hossain and Okui, 2013). Like these other methods, our procedure does not require the assumption of risk-neutrality, and only 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 procedure that maps their reported beliefs into the optimal weights according to the 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.

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 7% of weights did not correspond to the suggested optimal.[16]

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. Without additional assumptions, see Section 4.2, calculating the optimal weight requires knowledge of the probabilities of the two payoff relevant states: whether teammate 1 is top and teammate 2 is bottom, and vice-versa. In Wave 1 we assumed independence between beliefs about performance of the teammates, in order to calculate the probabilities of these states.

In Wave 2 beliefs were additionally elicited about the probabilities of all four possible states: both top, both bottom, and teammate 1 top and teammate 2 bottom (and vice-versa). Subjects had full freedom to re-allocate these probabilities to the four relevant states as they saw fit. Screenshots of the procedure can be seen in Figure 1 (and in Online Appendix Section 7 for Wave 1). Reassuringly, 90% of the time subjects chose not to alter beliefs in the four states, that is they followed the independence assumption.[17] In Online Appendix Section 1 we show beliefs are nearly identical across the two

[16]Results are not affected excluding these observations. Note that theoretically there are different combinations of beliefs (in particular sharing the same ratio) that lead to the same optimal weight. It is thus possible that subjects can arrive at the optimal weight, but intentionally report different combinations of beliefs to deceive the experimenter. We do not find this likely.

[17]Independence fails to hold after feedback, which create dependencies between beliefs about performance of the two teammates. For the 10% that reported beliefs that were inconsistent with the independence assumption, the average difference in the belief reported was less than one percentage point. Results are robust to excluding these observations. Piloting suggested it was not intuitive for subjects to initially think about the probabilities of these four states. For this

waves.

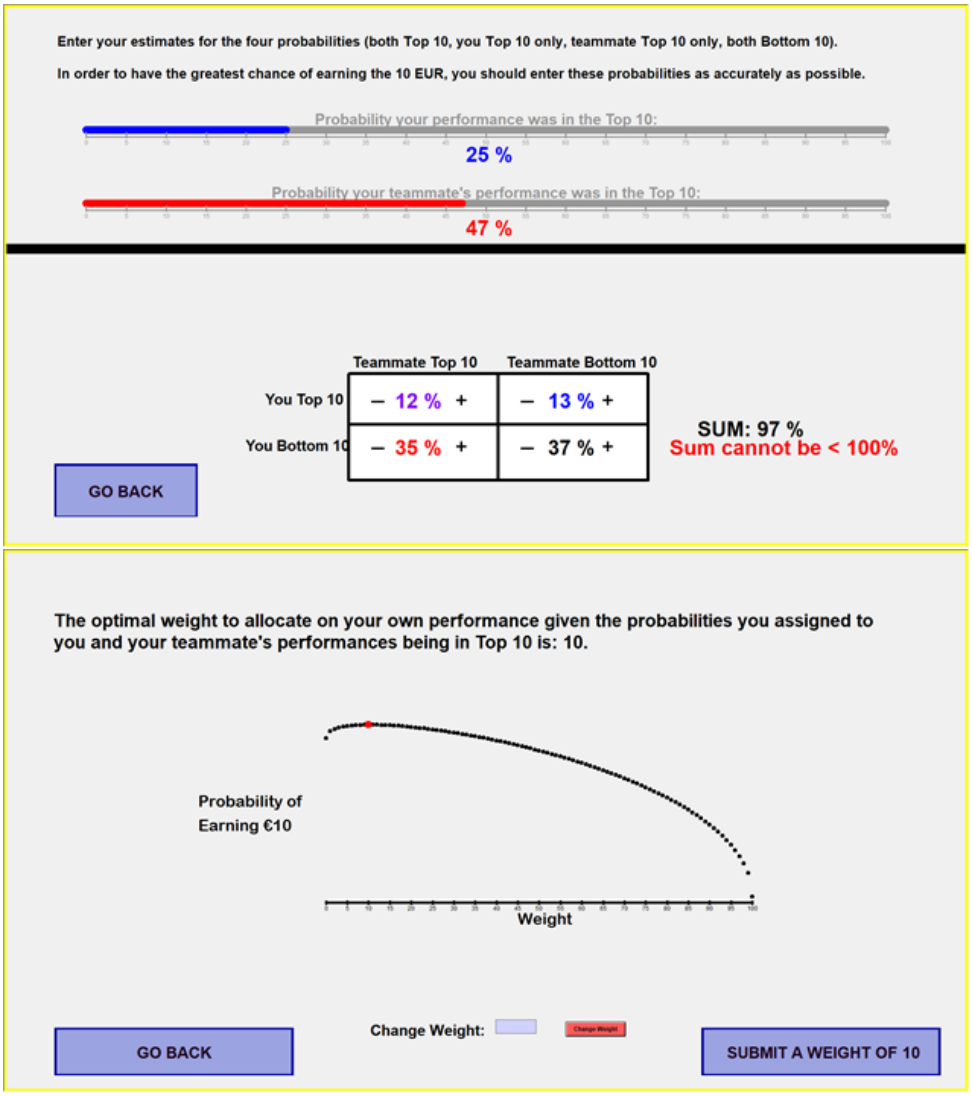


Figure 1: Screenshot of the mapping from chosen weight to probability of winning €10 which was calculated for every subject, conditional on the beliefs they entered.

3.5 Feedback

Once their weight was submitted subjects received feedback in the form of binary signals from a "Team Evaluator", represented as a cartoon figure. Signals are independent across time t and are positive (p) with probability $\Phi_{S_1S_2}$, otherwise they are negative (n). We denote them by $s_t = (p, n; \Phi_{S_1S_2})$.

Positive or negative team feedback corresponded in the experiment 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

reason we first asked about the probability of teammate 1 and 2 being in the top half.

teammates in Part 1 of the experiment and did not depend either on the beliefs reported by subjects nor the previous weights submitted. This ensured that subjects did not have incentives to "experiment" with their chosen 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. 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.[18] Before the start of Part 2, subjects had to answer five control questions that were aimed at ensuring their 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 answers provided were correct.

3.6 Part 3

In Wave 2, at the end of Part 2, we asked subjects their maximum willingness to pay (WTP) to switch their teammate 2 for Part 3, i.e. be randomly re-matched with a new teammate 2. Our interest in WTP stems from understanding the consequences of biases in attribution for decisions to change one's environment.

Part 3 otherwise was 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 to switch their teammate. The lottery would then choose a random price in the [€0, €5] interval and subjects would switch their teammate if their maximum WTP was above the chosen price and keep their teammate if this maximum is below that price. Our focus is on differences in WTP across Main and Control, this discussion is found in Section 6.2.4

4 Theory

4.1 Preliminaries

We first setup the theoretical framework which follows from the experimental design. An individual faces an uncertain environment with two sources of uncertainty: (i) the ability of teammate 1 (own ability in Main) (ii) the ability of teammate 2. Following the experiment, our interests are in the discrete 2×2 state space of the ability of both teammates. Teammate 1's unknown ability is given by $S_1 \in \{B, T\}$, corresponding to either low ability (bottom half of performance distribution) or

[18] 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.

high ability (top half). The unknown fundamental of interest $S_2 \in \{B, T\}$ is defined analogously, from the experiment this will correspond to whether teammate 2 is in the bottom half or top half of performances respectively. This leads to the four relevant states:

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

At any finite point in time t, the individual holds beliefs about the probability that the ability of teammate 1 and teammate 2 are T, given by b_t^1 and b_t^2 respectively.

As in the experiment, at each time period t, individuals take an action, by choosing how much to weight the performance of teammate 1 relative to teammate 2, ω_t. Monetary payoffs at time t, 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.

$$\Pi^t(\omega_t, S_1, S_2) = \begin{cases} P & \text{if } TT \\ (P, 0; \sqrt{\omega_t}) & \text{if } TB \\ (P, 0; \sqrt{1-\omega_t}) & \text{if } BT \\ 0 & \text{if } BB \end{cases} \tag{1}$$

4.2 Optimal weight

We now assume that individuals are subjective expected utility maximizers, with strictly increasing utility function $u(\cdot)$. Individuals form subjective beliefs about the probabilities that teammate 1 and 2 are in the top half. Thus, agents have priors b_0^1 and b_0^2 about the probability that $S_1 = T$ and $S_2 = T$ at time $t = 0$ respectively.

Denote beliefs about the four states at time t by: $b_t^{S_1 S_2}$. Let beliefs about all four states be a 4×1 vector, $\mathbf{b}_t$. At time $t = 0$, before receiving any feedback, the prior beliefs b_0^1 and b_0^2 are independent. Thus $b_0^{TT} = b_0^1 \cdot b_0^2$, 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, $u(\cdot)$, and hence is independent of risk preferences. Unless there is certainty, extreme weights are never optimal. 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. Thus, biases in beliefs regarding teammate 1 and 2 will be most costly when they are in opposing directions, e.g. an upward bias for teammate 1 and a downward bias for teammate 2. This will turn out to be critical for the interpretation of the results.

We now pause to note a few things. First, given the functional form of this expected utility function, $Q(\cdot)$, the optimum in Equation 4 is guaranteed to exist, and is unique for any beliefs except for the extreme case when $b_t^{TB} = b_t^{BT} = 0$.[19] 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. In our setup, the feedback that our agents receive is independent of their weighting decisions, which precludes this type of self-defeating learning.[20]

4.3 Belief Updating

We first examine the Bayesian benchmark to study how beliefs evolve for the four states, and hence how beliefs about being in the top half evolve. Note that $b_t^1 = b_t^{TT} + b_t^{TB}$. Following the experiment, from now on we also make explicit the assumption that: $1 > \Phi_{TT} = 1 - \Phi_{BB} > 0.5 = \Phi_{TB} = \Phi_{BT}$.

A Bayesian will update beliefs about teammate 1 being in the top half given either positive (p) or negative (n) signal respectively as follows:[21]

[19] 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 $b_t^{TB} + b_t^{BT} \leq 1$.

[20] 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 $b_0^1 = 1$. 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 b_0^1 and teammate 2 b_0^2. Then one can see that: $Q_{b_0^1}(\cdot) > 0$, $Q_{b_0^2}(\cdot) > 0$, $Q_{\omega b_0^2}(\cdot) < 0$, and $Q_{\omega b_0^1}(\cdot) > 0$. In other words, expected utility is increasing in expected ability of teammate 1 and 2, b_0^1 and b_0^2 respectively, and the optimal weight $\omega*$ is decreasing in the expected ability of teammate 2 b_0^2 and increasing in expected ability of teammate 1 b_0^1. While overconfidence and potentially biased updating in our context reduce expected payoffs, learning cannot be self-defeating in the sense of Heidhues et al. (2018).

[21] To derive this equation note (taking the case of a positive signal) that the probability of $s_t = p$ 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 a signal $s_t = p$ is $\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}$.

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

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

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

4.4 Self-Attribution Bias

Recalling the literature on attribution biases more generally, psychologists such as Heider drew important distinctions between attributions from stable qualities such as ability, versus unstable qualities such as randomness or luck. Our theoretical focus, following self-attribution bias, is on mis-attribution regarding own ability 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. For biased attribution to arise there must be some benefit to holding positive beliefs about one of these qualities. We focus on the case where the decision maker herself is teammate 1, corresponding to the Main part of the experiment. Thus, the driver of biased information processing comes from the possibility that decision makers benefit from inflated beliefs about their own ability. We are agnostic over the precise source of these benefits, among the possibilities outlined in Section 2.

In Appendix A we set forth a more general framework of how these potential benefits may interact with costs of overconfidence to motivate models of biased attribution. Here we focus on making primary distinctions in types of mis-attribution, but we will refer to this deeper framework as needed. In our context there are three possible sources to attribute feedback to: (1) performance of teammate 1 (self); (2) performance of teammate 2; (3) noise. Noise is present since signals are not perfectly informative about the states of the world, i.e. $\Phi_{S_1,S_2} \in (0,1)$. Mis-attribution will relate to inflating beliefs about factor (1) at the expense of factors (2) and (3).

With two dimensions of uncertainty, we are uniquely able to present a test of two models of quasi-Bayesian motivated attribution errors. A Bayesian attributes proportionately among these three sources. In the first model, noisy attribution bias (NAB), the agent processes information about other factors (teammate 2) accurately, 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 updates about (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 makes correct inferences about (3).

The psychology literature suggests that one should expect that the target of this mis-attribution

is more likely to be the other teammate (2) rather than noise (3), which corresponds to the model of FAB. In fact, in our micro-founded approach in Appendix A, which motivates these more reduced form biased updating models, the resulting model corresponds directly to our specification of FAB. A key reason for this is a consistency condition, which is required when evaluating the material costs of mis-stated beliefs in the framework.

Finally, we are able to highlight important consequences that may result from differences in these biases: namely that with FAB, individuals update in a biased but 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.[22]

4.4.1 Updating with Noisy Attribution Bias

With NAB, individuals over-attribute positive feedback to their own performance, and 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, such as Möbius et al. (2014) and our earlier references. Since we consider the additional dimension of uncertainty, the ability of teammate 2, we must additionally specify that NAB predicts that individuals update using Bayes' rule regarding the performance of this 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. Formally, they over-interpret the strength of the signal by a factor of γ_p, where $\gamma_p \geq 1$ in the case of a positive signal, and $\gamma_n \geq 1$ in the case of a negative signal. Our specification of the bias is thus similar to the biased updating model of Gervais and Odean (2001). Going more in depth, it is equivalent to one where the individual mis-interprets the strength of the signal in the states of the world consistent with their performance being in the top, i.e. TT and TB. However, updating about teammate 2's performance occurs as if the strength of the signal were correctly interpreted in all states, i.e. updating about teammate 2 is Bayesian.[23]

Thus, regarding own performance, biased updating in response to positive and negative feedback through NAB results in upward biased beliefs:

$$\left[b_{t+1}^{1,NAB}|s_t = p\right] = \frac{\gamma_p \left[\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}\right]}{\gamma_p \left[\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}\right] + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}} \geq \left[b_{t+1}^{1,BAYES}|s_t = p\right], \quad (7)$$

[22] Astute readers will note that NAB can also alter the final assessments of the other fundamental, indirectly through its affect on altering the interpretation of future signal likelihoods. As our empirical framework always incorporates prior beliefs, this will not alter our updating hypotheses.

[23] Our focus on beliefs abut teammates' being in the top is intentional: these are the most salient, and as noted, 90% of the time subjects enter these beliefs without further modification of the four states. We could have modeled biased updating through a mis-interpretation of the signal strength from any combination of the four states. However we don't see such directions providing additional insights beyond our specification. Moreover we find our formulation more natural if subjects form beliefs over the probabilities most relevant to their own ego, as opposed to the four underlying states.

$$\left[b_{t+1}^{1,NAB}|s_t=n\right]=\frac{\gamma_n\left[(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{TB})b_t^{TB}\right]}{\gamma_n\left[(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{TB})b_t^{TB}\right]+(1-\Phi_{BT})b_t^{BT}+(1-\Phi_{BB})b_t^{BB}} \geq \left[b_{t+1}^{1,BAYES}|s_t=n\right]. \tag{8}$$

With NAB, updating about their own performance exhibits positive asymmetry - 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.

4.4.2 Updating with Fundamental Attribution Bias

With FAB, individuals over-attribute positive feedback to their own performance, *at the expense* of the other source of uncertainty, i.e. 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.

$$\left[b_{t+1}^{1,FAB}|s_t=p\right]=\left[b_{t+1}^{1,NAB}|s_t=p\right]= \frac{\gamma_p\left[\Phi_{TT}b_t^{TT}+\Phi_{TB}b_t^{TB}\right]}{\gamma_p\left[\Phi_{TT}b_t^{TT}+\Phi_{TB}b_t^{TB}\right]+\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}} \geq \left[b_{t+1}^{1,BAYES}|s_t=p\right], \tag{9}$$

$$\left[b_{t+1}^{1,FAB}|s_t=n\right]=\left[b_{t+1}^{1,NAB}|s_t=n\right]= \frac{\gamma_n\left[(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{TB})b_t^{TB}\right]}{\gamma_n\left[(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{TB})b_t^{TB}\right]+(1-\Phi_{BT})b_t^{BT}+(1-\Phi_{BB})b_t^{BB}} \geq \left[b_{t+1}^{1,BAYES}|s_t=n\right]. \tag{10}$$

The key difference is that with FAB the individual updates in a biased but consistent manner across themselves and their teammate, unlike with NAB, where they update as a standard Bayesian with respect to their teammate. The fact that our specification of self-attribution bias implies a distortion of the underlying states also suggests that biased attribution should satisfy a consistency condition when updating about the probability that teammate 2 is in the top half. Specifically in response to positive and negative signals respectively it implies:

$$\left[\hat{b}_{t+1}^{2}|s_t=p\right]=\frac{\gamma_p\Phi_{TT}b_t^{TT}+\Phi_{BT}b_t^{BT}}{\gamma_p\left[\Phi_{TT}b_t^{TT}+\Phi_{TB}b_t^{TB}\right]+\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}} \tag{11}$$

$$\left[\hat{b}_{t+1}^{2}|s_t=n\right]=\frac{\gamma_n(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{BT})b_t^{BT}}{\gamma_n\left[(1-\Phi_{TT})b_t^{TT}+(1-\Phi_{TB})b_t^{TB}\right]+(1-\Phi_{BT})b_t^{BT}+(1-\Phi_{BB})b_t^{BB}} \tag{12}$$

Whether the biased posterior $\left[b_{t+1}^{2,FAB}|s_t=s\right]$ is smaller or larger than the Bayesian $\left[b_{t+1}^{2,BAYES}|s_t=s\right]$ depends on the signal likelihoods $\Phi_{S_1S_2}$. Given our assumptions, both $\Phi_{TT}\Phi_{BB}-\Phi_{TB}\Phi_{BT}=(1-\Phi_{TT})(1-\Phi_{BB})-(1-\Phi_{TB})(1-\Phi_{BT})\leq 0$. As shown in Appendix B this guarantees that

$\left[b_{t+1}^{2,FAB}|s_t = s\right] \leq \left[b_{t+1}^{2,BAYES}|s_t = s\right]$.[24]

Thus, FAB implies that when considering their teammate's performance they update asymmetrically in the *negative* direction, i.e. they under (over)-weight positive (negative) signals. Regarding the material consequences, we can also note that because the optimal weight depends on relative performance beliefs across the two teammates, FAB implies a greater expected financial penalty, due to the additional downward bias on teammate 2 beliefs, leading to a more extreme weight.

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

While our main focus is on updating beliefs we also discuss belief formation and present hypotheses relating to overconfidence biases, which presents a litmus test for whether subjects find the IQ task ego-relevant. 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. Following, Dubra and Benoît (2011), 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: formed and updated according to the properties of Bayes' rule.

Our first null hypothesis of interest concerns whether there is overconfidence in the Main treatment for teammate 1. 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:

[24] In the experiment both of these conditions are equivalent to $0.9 \cdot 0.1 - 0.5 \cdot 0.5 < 0$. In addition, without making any assumptions about updating, the condition ensuring $\left[b_{t+1}^{2,FAB}|s_t = s\right] \leq \left[b_{t+1}^{2,BAYES}|s_t = s\right]$ is only violated in 2% of updating cases, also shown in Appendix B. Excluding these observations does not alter the results.

Hypothesis 1*:

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

where C indicates the Control treatment and the elicited belief is about a third party. Examining these two hypotheses together presents a 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, see Benjamin (2019) for a summary, have shown that there are important deviations from Bayes' rule in belief updating. Due to expected deviations from Bayes' rule, as in Section 5.1 we again 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. The implication is that *the optimal weight will be constant*: that is, it will be non-responsive to feedback in our context. As 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:

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

Analogously for teammate 2:

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

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); see Benjamin (2019) for additional references. Bayes' rule can be written in the following form, considering binary signals, $s_t = k \in \{p, n\}$, for positive and negative respectively, 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 $s_t = k \in \{p, n\}$ 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 derived similarly, noting the difference between NAB and FAB. 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:

$$\frac{\gamma_p \left[0.9b_t^{TT} + 0.5b_t^{TB}\right]}{b_t^{TT} + b_t^{TB}},$$

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, $\hat{LR}_t^1(p)$, is thus:

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

Similarly, the ratio, $\hat{LR}_t^1(n)$, is:[25]

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

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

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

The empirical model nests this Bayesian benchmark as follows,

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

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

[25] We note that there is an implicit upper bound on γ_n as this equation is ≤ 1. The reason is that we must assume that a negative signal is in fact perceived as negative information. If γ_n were implausibly large, the interpretation of this would be that biased individuals actually perceive negative signals as indicating a greater likelihood of performing in the top half. Within the context of our deeper theoretical model in Appendix A, we interpret this as a restriction on the shape of the mental costs of distorting γ_n.

experimental subject.[26]

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 $\hat{LR}_t^1(p) \geq LR_t^1(p)$ and $\hat{LR}_t^1(n) \geq LR_t^1(n)$. Bayesian posteriors result in a weight of $\beta_1 = 1$ or $\beta_0 = 1$ on $LR_t^1(p)$ or $LR_t^1(n)$ 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.[27] In other 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). Since our theories of attribution bias do not alter predictions of δ, we remain agnostic over these values. 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):
$\beta_1^{1,M} > 1;\ \ \beta_0^{1,M} < 1$
$\beta_1^{2,M} = 1;\ \ \beta_0^{2,M} = 1$

Hypothesis 3*:

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

[26] Regarding the robustness of this framework, Online Appendix Section 2 presents simulated updating data, and finds that adding mean-zero noise to posteriors does not generate meaningful asymmetry, but can generate conservatism and $\delta < 1$. Our main focus in Section 6 is on asymmetry, and moreover, on the differences in updating across Main and Control treatments.

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

$\beta_1^{2,M} = \beta_1^{2,C};\ \ \beta_0^{2,M} = \beta_0^{2,C}$

Hypothesis 4:

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

Hypothesis 4*:

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

6 Results

6.1 Initial Beliefs

First we investigate whether prior beliefs are well calibrated in the first round. Prior beliefs in Main and Control treatments for both teammates are presented in Figure 2. 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 also 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 intriguingly 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.[28]

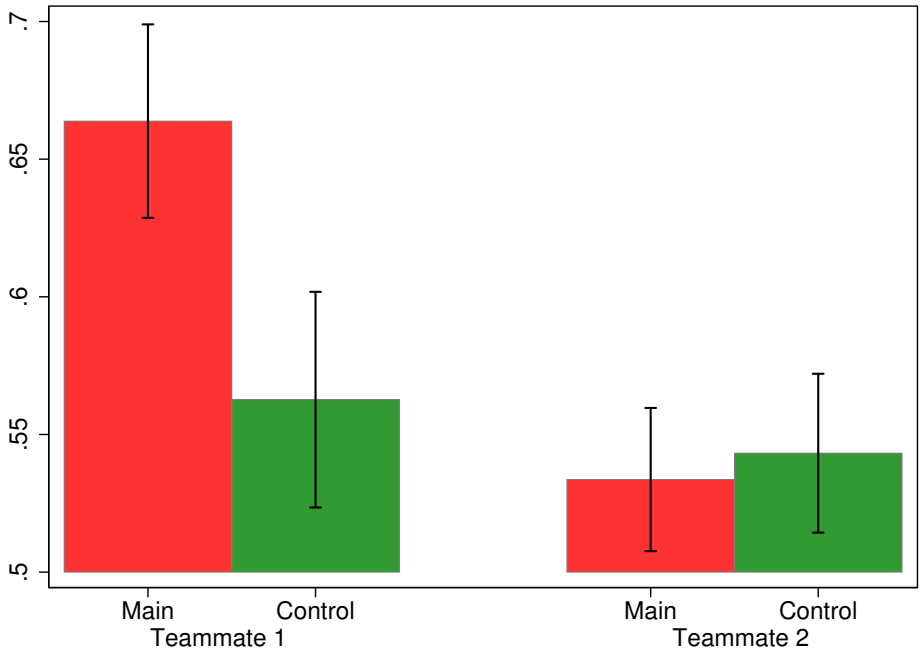
Thus, following Hypothesis 1*, we compare beliefs across the two treatments, Main and Control. In fact we can reject equality of mean prior beliefs across the two settings at the 1% level (Wilcoxon rank-sum test p-value: 0.0005). This provides robust evidence that what we are observing in the Main treatment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

[28]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.

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 (Wilcoxon rank-sum p-value is 0.5723). Additionally Wilcoxon signed rank tests also reject the hypotheses that these beliefs are equal to 50% at the 1% level (p-values 0.001 and 0.002 respectively), again pointing to an upward bias in beliefs that cannot be explained by overconfidence.

We also note that our hard-easy manipulation was successful. More details are provided in Online Appendix Section 3, however individuals rate themselves in the top half with 72% probability when the test was easy, and 62% when the test was hard. While not our main focus, we find evidence that men are more overconfident than women. Further details are provided in Online Appendix Section 4. We defer discussions about gender differences to our analysis on belief updating, where we control for subjects' priors.

Figure 2: Prior Beliefs by Treatment



For teammate 1: Main, Belief about own performance; Control, Belief about other teammate 1's performance. For teammate 2: Belief about other teammate 2's performance. 95% Confidence intervals.

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, utilizing a matching strategy. While our main focus is on beliefs, we also present an analysis of the resulting weights in Appendix C as well as WTP to be matched to a new teammate 2 in Section 6.2.4.

6.2.1 Structural Framework

The primary analysis follows the framework outlined in Section 5.2.1. Table 2 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 sample includes all updates from both waves, in Part 2 and 3.[29]

Table 2: 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 2. From now on we focus only on our hypotheses comparing updating to the Control, rather than the Bayesian benchmark.

From Table 2 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 performance. 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.

[29] Samples excluding Part 3 are presented in Online Appendix Section 5, with similar results. We follow common sampling restrictions in the literature: excluding boundary observations and wrong direction updates. With two-dimensional uncertainty, we classify a wrong direction update as updating at least one belief in the wrong direction, without compensating by adjusting the other belief in the correct direction. More details provided in Online Appendix Section 5.

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) asymmetric updating in Main compared with Control.

Table 3: 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.

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 3 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.

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). 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 2 column 1 is significantly greater than β_1 in Table 3 column 1 (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.

To explain these results we consider variations of our model of optimal distorted information processing in Appendix A. In that model a subconscious process selects the optimal degree of biased information processing about one's own performance, to trade-off the costs and benefits to overconfidence. In light of our results, we relax the Bayesian framework even further to allow individuals to also update in a biased way about teammate 2. This provides a richer toolset for the subconscious process to use.

In fact, we show that positive asymmetry for *both teammates* under this extended model can be optimal. The intuition for this result is that while updating in a positive asymmetric way about teammate 2 will result in slightly lower beliefs about own performance, there is a countervailing effect which lowers the material costs of biased information processing. The reason for the latter is that the resulting optimal weight will be more moderate, since the individual will assign a lower weight to being in the top half relative to their teammate. This counteracts the damaging effects of overconfidence, by leading to a weight which generates a higher probability of earning the €10 in the experiment. We discuss this model in more detail in Appendix A.1.

Of note is that there are a few candidate alternative explanations for these results. We discuss two more prominent ones: first, that anchoring causes individuals to update similarly about teammate 2, and second that subjects selectively ignore negative signals. We find evidence suggesting that these two explanations cannot explain the patterns in our data, namely as raw absolute and percentage updates are not positively correlated across teammate 1 and 2 in our Main treatment, and subjects in our Main treatment respond to negative signals at equivalent rates to those in the Control treatment. We address these alternative explanations in more detail in Online Appendix Section 6.

Thus to summarize the results of our structural framework, individuals are positively asymmetrically biased about both team members in our Main treatment, but not in Control. Such biased updating cannot be explained by our initial framework of NAB or FAB, but can be rationalized in a broader framework where individuals can optimally update in a biased way about themselves and their teammate, in order to maximize the trade-offs between the costs and benefits of overconfidence. The flexibility to update in a biased way about one's teammate enables subjects to potentially indulge in even greater overconfident beliefs, while mitigating the material costs through inflated beliefs about their teammates.[30]

[30] While our model implies relatively sophisticated sub-conscious behavior, we note that the intuition for why positive bias about both teammates leads to more moderate weights is relatively straightforward.

Lastly, we turn to briefly examining potential differences in updating by gender. Online Appendix Section 4 presents these tables. 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 the finding of asymmetry is not statistically significant for females only. These patterns are true both for updating about teammate 1 and teammate 2.

6.2.2 Non-Parametric Inference

The previous section provides significant evidence of differential updating between the Main and Control treatments, along with a discussion of the potential mechanisms. However, the structural framework makes relatively rigid assumptions about the precise form of updating. Here we examine whether updating is different between our two treatments outside of this framework. This has the advantage of not requiring any assumptions about how subjects update. However, this also comes at a cost in that we will be limited in our ability to identify or corroborate the mechanisms identified in the previous section.

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, relative to the Bayesian benchmark. For this we consider rounds 1 through 5 in Part 2 common to all subjects, but not Part 3.[31] 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 2 and 3 for comparability.

For this first look, in Appendix Figures D.1 and D.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 D.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.

For the second look, Figure D.4 shows 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,

[31] 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.

posteriors in Main show an upward bias relative to Control, for both positive and negative signals, consistent with the patterns in Table 2. 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. We now turn to a matching strategy which will control for differences in initial priors across Main and Control, while allowing us to make aggregate inferences about updating behavior.

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.[32] 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 randomly chosen teammate (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 total number of negative signals received over the four rounds of feedback.[33]

Table 4 presents the results of this exercise reporting average treatment effects (ATE). 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 randomly chosen teammate 1, conditional on having the same priors and facing the *same sequence* of signals. This indicates that information processing differs across the two treatments.

Table 4: Main vs Control: Belief Teammate 1 Top

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

Analysis uses nearest neighbor matching, with replacement. 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.

[32] 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.

[33] Priors of matched neighbors must be within 3 percentage points, i.e. a caliper of 0.03. The results (available upon request) are consistent for other calipers.

Table 5: Main vs Control: Belief Teammate 1 Top by Number of Received Negative Signals

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 4 −
ATE	−0.016	0.105	0.134***	−0.025	0.185**
	(0.073)	(0.083)	(0.047)	(0.087)	(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).

The empirical framework suggests this difference in updating is driven primarily by under responsiveness to negative signals. To investigate this, Table 5 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 a balanced sequence of 2 positive and 2 negative signals. Overall these patterns are supportive of the structural results.

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 6 and 7. 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 points greater in Main relative to Control.

Table 6: Main vs Control: Belief Teammate 2 Top

	(1) 1 Neighbor	(2) 2 Neighbors
ATE	0.045	0.040
	(0.035)	(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 7: Main vs Control: Belief Teammate 2 Top by Number of Received Negative Signals

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 4 −
ATE	−0.016	0.075	0.031	−0.004	0.139**
	(0.101)	(0.099)	(0.077)	(0.096)	(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).

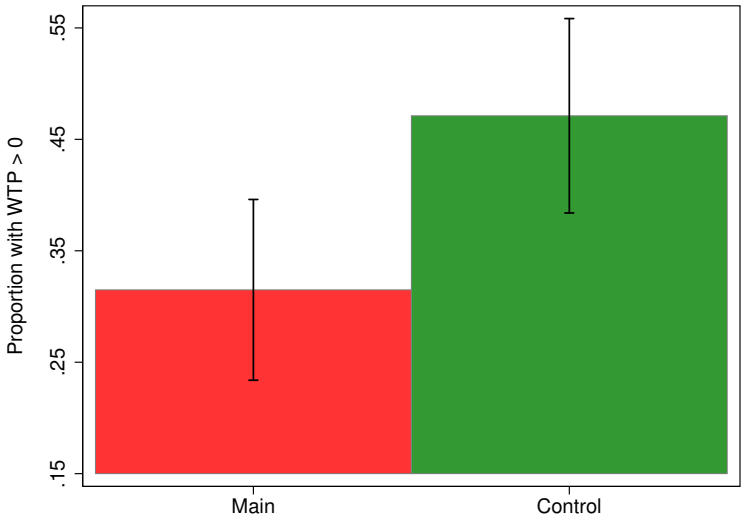
6.2.4 Willingness to Change Teammates

Finally, we briefly discuss the implications of observed biased updating, in the context of future decision making. Recall that in Wave 2 we measured the willingness of subjects to replace teammate 2 with a new (randomly selected) teammate, by submitting a willingness to pay (WTP) between 0 and 5€. Here we focus on the binary decision of whether a subject is willing to change teammates (i.e. WTP$>$ 0). Ex-ante, the proportion of those willing to switch teammates should be the same in Main and Control treatments. This holds true, even with the overconfident prior beliefs observed in Main, since, before feedback, the decision to change teammates depends only on the belief about the performance of teammate 2, see Appendix E.[34]

Given the patterns of biased updating we observe in our Main treatment, subjects end up with more positive performance beliefs about teammate 2. This lowers the proportion of subjects in Main who should be willing to pay to switch teammates, as Appendix E confirms given actual subject beliefs after 4 rounds of feedback. We also confirm this outcome in our actual WTP data. Figure 3 presents the proportion of subjects who submit a WTP strictly greater than zero, by Main and Control treatments. 31% of Main subjects and 47% of Control subjects were willing to pay to change teammates, a difference significant at the 5% level (Ranksum p-value 0.0155). As expected, as a result of biased updating about teammate 2, subjects in Main treatment are 34% less likely to want to change teammates than their Control counterparts. This confirms that the biased updating patterns we observed translate into actual differences in future decision making. Moreover, it suggests that subjects are sufficiently confident about their reported beliefs that they act on them in a context which falls outside of the purview of the elicitation procedure.

[34]Note that initial beliefs about own performance affect the value of one's WTP, not whether it is positive or negative. Higher performance beliefs lead to a lower value of switching teammates, since the weight allows one to hedge against having a lower performing teammate. Correspondingly, we do find that WTP is smaller in Main vs Control, among those submitting a positive WTP (intensive margin), though it is not significant at conventional levels.

Figure 3: Willingness to pay



Proportion of subjects who submitted strictly positive WTP to change teammate 2. Wave 2 only ($N = 231$). 95% confidence intervals shown.

7 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 to attribute positive feedback to internal qualities about themselves, and negative feedback to salient external factors. We took this theory and quantified it, placing it within a quasi-Bayesian updating framework where individuals face two dimensions of uncertainty. In the context of a naturally framed lab experiment with two person teams, we examined how individuals attributed feedback between themselves, their teammate, and noise. 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. With regard to FAB, we provided additional micro-foundations for the optimal level of biased information processing, trading off the benefits and costs of holding overconfident beliefs.

In our results we document significant evidence of overconfidence and subsequent biased patterns 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, their updating about their teammate follows similar patterns, over-weighting positive relative to negative feedback. Notably, in our Control treatment, individuals update symmetrically when receiving positive or negative feedback.

Neither FAB or NAB fit the patterns of positive asymmetry for both teammates, i.e. both models are rejected. However, as a result of this positive bias for teammate 2, the submitted weight is more moderate, and the resulting losses from overconfidence are lower than they would otherwise be, absent no bias or negative asymmetry in updating about teammate 2. As such, material losses from overconfidence in the experiment are mitigated. A further extension of our micro-founded model of FAB, where we allow biased updating about both teammates, confirms that such a strategy is in fact

optimal. Importantly, this extended model provides a possible explanation for why we observe strong positive asymmetry, while some other studies have not - the possibility of biased updating for another dimension of uncertainty permits even stronger self-serving beliefs, by providing a tool which in certain contexts can mitigate the material costs of overconfidence.

Our structural results are consistent with additional non-parametric tests. Our estimates suggest 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 when receiving all negative signals is that they end up more optimistic about their teammate's performance as well.

Our results and theoretical insights provide important contributions to our understanding of 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 spill over to their beliefs about the performance of others. One clear implication of this is that individuals will thus make inefficient decisions not only because of their own overconfidence, but also relating to their biased beliefs about other states of the world.

Specifically, in contexts where individuals interact repeatedly in a fixed environment with another source of uncertainty, positive asymmetry about this source can mitigate the consequences of overconfidence. Thus, while Heidhues et al. (2018) show negative consequences from self-defeating learning, our results suggest that being additionally biased about the fundamental can lower these costs. On the other hand, when individuals have opportunities to change environments, as in Hestermann and Le Yaouanq (2018), our results suggest that overconfident individuals may be more likely to stay in lower quality environments, due to positively biased beliefs about the unknown fundamental. This implication is borne out in our data: individuals in our Main treatment are significantly less likely to demand to change environments. This contrasts with the results of Hestermann and Le Yaouanq (2018), who showed that underconfident individuals were more likely to be trapped in low quality environments. As most real world informational environments are complex and involve more than one dimension of uncertainty, our theory and empirical results suggest important new insights about how information is processed in such settings. A key takeaway is that biases which enable self-serving beliefs do not exist in a vacuum, and can lead to distorted perceptions about the broader world.

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.

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

_ and Jean Tirole, "Over My Dead Body: Bargaining and the Price of Dignity," *American Economic Review*, 2009, *99* (2), 459–465.

_ and _ , "Identity, morals, and taboos: Beliefs as assets," *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

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.

Bracha, Anat and Donald J. Brown, "Affective decision making: A theory of optimism bias," *Games and Economic Behavior*, 5 2012, *75* (1), 67–80.

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

Burks, S. V., J. P. Carpenter, L. Goette, and a. Rustichini, "Overconfidence and Social Signalling," *The Review of Economic Studies*, 1 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*, 6 2019, *22* (2), 369–395.

_ , "Testing models of belief bias: An experiment," *Games and Economic Behavior*, 1 2019, *113*, 549–565.

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.

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.

_ **and John L. Michela**, "Attribution Theory and Research," *Annual Review of Psychology*, 1 1980, *31* (1), 457–501.

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*, 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, Carmen and Alfonso Rosa-García, "Gender bias in risk aversion: evidence from multiple choice exams," *mimeo*, 2011, (39987).

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, 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.

Pekrun, Reinhard and Herbert W. Marsh, "Weiners attribution theory: Indispensablebut 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.

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.

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 Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices," *Journal of Labor Economics*, 1 2014, *32* (1), 161–198.

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *American Economic Review*, 2019.

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 Model of Optimal Information Distortion

In this section we provide a micro-foundation for FAB. Specifically we follow Brunnermeier and Parker (2005) by assuming that agents engage in a subconscious optimization problem which selects the optimal $\gamma_s \in \mathbb{R}_+$ at the moment the individual processes new information, trading off the benefits from overconfidence against the costs. While updating beliefs over time is a dynamic problem, we assume a static model of updating. We wish to avoid the additional complexity involved in a dynamic model of optimally biased updating, but additionally, our focus here is on the short-run.[35] Unlike Brunnermeier and Parker (2005) we relax the assumption of Bayesian updating, and assume that this optimization occurs directly over the updating process, through parameters γ_p, γ_n rather than beliefs b_{t+1}. The updating process is precisely that outlined in FAB, for Equations 9 through 12.

We introduce the possibility that individuals receive direct utility over the belief that they are in the top half, through a function $\alpha \cdot v(b_{t+1})$, where $v(\cdot)$ is a monotonically increasing, concave, twice continuously differentiable function. $\alpha \in \{0, 1\}$ indicates whether or not the individual benefits from holding overconfident beliefs. This can be thought of as a reduced form interpretation of the benefits to overconfidence, e.g. direct hedonic utility benefits, signalling to others, or benefits from motivation. Importantly, we assume that individuals do not derive any benefit from beliefs about others' ability, nor do they derive direct benefit from beliefs about the four states TT, TB, BT, BB. Of course, since $b_t^1 = b_t^{TT} + b_t^{TB}$, indirectly they can benefit from these beliefs.

We propose that a subconscious process trades off these benefits from overconfidence against the costs, which we posit to be material costs from inefficient decision making as well as mental costs of distorting the updating process. In the experiment, these material costs are the lower expected probability of earning $P = €10$. Following Bracha and Brown (2012), we assume a mental cost function $J(\gamma_s, 1)$ that is convex, strictly increasing in $|\gamma_s - 1|$, and is minimized at the Bayesian information processing parameter $\gamma_s = 1$.[36]

In the following we denote $\hat{b}_{t+1}$ as potentially biased beliefs, with b_{t+1} referring to the posteriors that would arise following Bayes rule. We first note that if subjects hold biased beliefs, they will submit a distorted weight in the experiment, which generates material costs from foregone income. These will relate to the chosen weight $\hat{\omega}^*_{t+1}$, given potentially biased posterior beliefs. Critically, the optimal weight depends on beliefs about two states b^{TB}_{t+1} and b^{BT}_{t+1}. Given the form of the updating bias for both NAB and FAB for updating about own ability, this implies an over-weighting of the likelihood of state TB by γ_s. Yet while FAB embodies the same likelihood error underlying the updating decision about teammate 2, NAB prescribes no such over-weighting when updating about teammate 2. Hence NAB generates a clear logical contradiction in the current framework.

As a result, in what follows the only consistent way forward is to follow FAB, and impose consistency across updating about teammate 1 and 2, which leads to the (biased) optimal weight $\hat{\omega}^*_{t+1}$.

[35] Long run models of belief distortion are studied by Heidhues et al. (2018) and Möbius et al. (2014).

[36] Following Bracha and Brown (2012) we further assume that $\lim_{\gamma_s \to \{\infty\}} J'(\gamma_s, 1) \to \infty$. Intuitively, absent monetary incentives the model would always predict extreme overconfidence, which seems implausible. Justification for such a cost function are discussed in Bracha and Brown (2012). Finally, experimental evidence suggests that such mental costs are necessary if one wishes to take models of belief distortion seriously, see Coutts (2019b).

From Equation 4, setting $\Phi_{BT} = \Phi_{TB} = 0.5$, we have[37]:

$$\hat{\omega}^*_{t+1} = \frac{1}{1 + \left(\frac{\hat{b}^{BT}_{t+1}}{\hat{b}^{TB}_{t+1}}\right)^2} = \frac{1}{1 + \left(\frac{b^{BT}_t}{\gamma_s b^{TB}_t}\right)^2}. \tag{18}$$

We can now write the following, noting again that $s \in \{p, n\}$ refers to whether the individual received a positive or negative signal respectively.

$$\max_{\{\gamma_s\}} \Big\{ \alpha \cdot v(\hat{b}_{t+1}) + b^{TT}_{t+1} \cdot u(P) + b^{TB}_{t+1} \cdot \sqrt{\hat{\omega}^*_{t+1}} \cdot u(P) + b^{TB}_{t+1} \cdot (1 - \sqrt{\hat{\omega}^*_{t+1}}) \cdot u(0)$$
$$+ b^{BT}_{t+1} \cdot \sqrt{1 - \hat{\omega}^*_{t+1}} \cdot u(P) + b^{BT}_{t+1} \cdot (1 - \sqrt{1 - \hat{\omega}^*_{t+1}}) \cdot u(0) + b^{BB}_{t+1} \cdot u(0) - J(\gamma_s, 1) \Big\},$$

where $\hat{b}_{t+1}$ and $\hat{\omega}^*_{t+1}$ are given by the distorted updating process. Objective (Bayesian) posteriors are $b^{S_1 S_2}_{t+1}$ for the four states. An important simplifying assumption that we maintain is that the subconscious process optimizes myopically each period, taking priors as given. Thus while priors may also be biased, we assume the subconscious process is naive in this respect, and takes priors at face value.

When $\gamma_s = 1$ there is no bias in the weighting decision. Importantly, $\gamma_s \neq 1$ will introduce an inefficient distortion, and leads to a lower probability of earning $P > 0$.

Substituting biased beliefs and weights into the maximization, and substituting the values of Φ from the experiment, the first order condition of this expression is:

$$\alpha \cdot \frac{\partial v}{\partial \gamma_s} + b^{TB}_{t+1}\Big(u(P) - u(0)\Big) \frac{\left(b^{BT}_t\right)^2 \left(\frac{\gamma_s^2 \left(b^{TB}_t\right)^2}{\left(b^{BT}_t\right)^2 + \gamma_s^2 \left(b^{TB}_t\right)^2}\right)^{\frac{3}{2}}}{\gamma_s^3 \left(b^{TB}_t\right)^2}$$
$$- b^{BT}_{t+1}\Big(u(P) - u(0)\Big) \frac{\gamma_s \left(b^{TB}_t\right)^2 \left(\frac{\left(b^{BT}_t\right)^2}{\left(b^{BT}_t\right)^2 + \gamma_s^2 \left(b^{TB}_t\right)^2}\right)^{\frac{3}{2}}}{\left(b^{BT}_t\right)^2} - J'(\gamma_s, 1)$$

Simplifying, and setting $u(P) - u(0) = \Delta u$:

$$\frac{\alpha}{\Delta u} \cdot \frac{\partial v}{\partial \gamma_s} + \frac{\left(b^{TB}_t b^{BT}_t\right)^2}{\left(\left(b^{BT}_t\right)^2 + \gamma_s^2 \left(b^{TB}_t\right)^2\right)^{\frac{3}{2}}} \cdot (1 - \gamma_s) - \frac{J'(\gamma_s, 1)}{\Delta u}$$

[37] We note that, given the biased updating process, this is simplified from the following equation (analogously for a negative signal): $\frac{\hat{b}^{BT}_{t+1}}{\hat{b}^{TB}_{t+1}} = \frac{\frac{\Phi_{BT} b^{BT}_t}{\gamma_p [\Phi_{TT} b^{TT}_t + \Phi_{TB} b^{TB}_t] + \Phi_{BT} b^{BT}_t + \Phi_{BB} b^{BB}_t}}{\frac{\gamma_p \Phi_{TB} b^{TB}_t}{\gamma_p [\Phi_{TT} b^{TT}_t + \Phi_{TB} b^{TB}_t] + \Phi_{BT} b^{BT}_t + \Phi_{BB} b^{BB}_t}} = \frac{\Phi_{BT} b^{BT}_t}{\gamma_p \Phi_{TB} b^{TB}_t}$.

The first term is positive, while the second is positive for $\gamma_s < 1$, and negative for $\gamma_s > 1$. The third term is always negative. While the second term is always negative for $\gamma_s > 1$, it is decreasing until it reaches an absolute minimum, then it increases, towards 0 as $\gamma_s \to \infty$. This could potentially lead to 2 solutions to the FOC (satisfying conditions for a maximum, though likely only one is the global max.) We avoid this issue by assuming properties of the mental cost function such that the overall the total (mental plus material) marginal costs are always strictly increasing in γ_s. This guarantees a unique solution. Existence is guaranteed by the intermediate value theorem.[38]

When $\alpha = 1$, i.e. there are benefits to overconfidence, $\gamma_s > 1$. For non-degenerate beliefs, the resulting biased updating leads to inflated posteriors about own performance, and deflated posteriors about the teammate's performance. Posteriors about the teammate's performance may not always be lower. In our setting they are, see Appendix B. Importantly when $\alpha = 0$, i.e. there are no benefits to overconfidence, the optimal $\gamma_p = \gamma_n = 1$, i.e. posteriors are Bayesian.

A.1 Biased updating about both Teammates

Here we extend the model by positing that individuals update their beliefs in a potentially biased way for *both* themselves and their teammate. We thus specify two bias parameters for each type of signal, one for self (teammate 1), and the other for teammate 2. We denote these parameters by $(\gamma_p^1, \gamma_p^2, \gamma_n^1, \gamma_n^2)$, or in short-form (γ_s^1, γ_s^2), $s \in \{p, n\}$. We also introduce the analogous mental cost function for the distortion of γ_s^2. Finally we also assume that updating across teammates must lead to consistent updating about each of the four states.

Under this formulation we note that resulting biased posterior beliefs for teammate 1 and 2 will be (showing the case for a positive signal):

$$\left[\hat{b}_{t+1}^1 | s_t = 1\right] = \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \tag{19}$$

$$\left[\hat{b}_{t+1}^2 | s_t = 1\right] = \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^2\Phi_{TB}b_t^{TB}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}}. \tag{20}$$

Optimal weights will be:[39]

$$\hat{\omega}_{t+1}^* = \frac{1}{1 + \left(\frac{\gamma_s^2 b_t^{BT}}{\gamma_s^1 b_t^{TB}}\right)^2} \tag{21}$$

[38] We note that at $\gamma_s = 1$ the FOC is positive when $\alpha = 1$. It is decreasing as γ_s increases (SOC is negative), and due to the properties of the mental cost function (convex, strictly increasing, with $J'(\gamma_s) \to \infty$ as $\gamma_s \to \infty$) there exists some γ_s^* where the FOC equals 0.

[39] As before, we note that (for the case of a positive signal):

$$\frac{\hat{b}_{t+1}^{BT}}{\hat{b}_{t+1}^{TB}} = \frac{\frac{\gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}}}{\frac{\gamma_p^1\Phi_{TB}b_t^{TB}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}}} = \frac{\gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\Phi_{TB}b_t^{TB}}.$$

Leading to the following optimization problem:

$$\max_{\{\gamma_s\}} \Big\{ \alpha \cdot v(\hat{b}^1_{t+1}) + b^{TT}_{t+1} \cdot u(P) + b^{TB}_{t+1} \cdot \sqrt{\hat{\omega}^*_{t+1}} \cdot u(P) + b^{TB}_{t+1} \cdot (1 - \sqrt{\hat{\omega}^*_{t+1}}) \cdot u(0) \\ + b^{BT}_{t+1} \cdot \sqrt{1 - \hat{\omega}^*_{t+1}} \cdot u(P) + b^{BT}_{t+1} \cdot (1 - \sqrt{1 - \hat{\omega}^*_{t+1}}) \cdot u(0) + b^{BB}_{t+1} \cdot u(0) \\ - J(\gamma^1_s, 1) - J(\gamma^2_s, 1) \Big\}$$

Substituting biased beliefs and weights into the maximization, and substituting the values of Φ from the experiment, the first order condition with respect to γ^1_s is (where $u(P) - u(0) = \Delta u$):

$$\frac{\alpha}{\Delta u} \cdot \frac{\partial v}{\partial \gamma^1_s}\left(\hat{b}^1_{t+1}\right) + \frac{\gamma^2_s \cdot \left(b^{TB}_{t+1} \cdot b^{BT}_{t+1}\right)^2}{\left(\left(\gamma^2_s b^{BT}_{t+1}\right)^2 + \left(\gamma^1_s b^{TB}_{t+1}\right)^2\right)^{\frac{3}{2}}} \cdot (\gamma^2_s - \gamma^1_s) - \frac{J'(\gamma^1_s, 1)}{\Delta u} \tag{22}$$

The first order condition with respect to γ^2_s is:

$$\frac{\alpha}{\Delta u} \cdot \frac{\partial v}{\partial \gamma^2_s}\left(\hat{b}^1_{t+1}\right) + \frac{\gamma^1_s \cdot \left(b^{TB}_{t+1} \cdot b^{BT}_{t+1}\right)^2}{\left(\left(\gamma^2_s b^{BT}_{t+1}\right)^2 + \left(\gamma^1_s b^{TB}_{t+1}\right)^2\right)^{\frac{3}{2}}} \cdot (\gamma^1_s - \gamma^2_s) - \frac{J'(\gamma^2_s, 1)}{\Delta u} \tag{23}$$

Again, we can note that if $\alpha = 0$, the optimal $\gamma^1_s = \gamma^2_s = 1$. When $\alpha = 1$, $\gamma^1_s > 1$, while the optimal γ^2_s may be less than, equal to, or greater than 1, though $\gamma^s_2 \leq \gamma^s_1$.[40] The arguments for $\gamma^1_s > 1$ are the same as in the previous model. The reason why γ^2_s cannot be pinned down is that while, given our assumptions, the first term is negative, the second term will in fact be positive for $\gamma^2_s < \gamma^1_s$ (not 1, as before). Thus there can be a benefit to updating in a biased way about teammate 2, which in fact counter-balances the biased updating about teammate 1, leading to a closer to optimal weighting decision.

A.1.1 Example

Here we denote an example taking on a specific functional form to illustrate the properties mentioned above. In particular we assume that $v(b^1_t) = \sqrt{b^1_t}$, $U(P) - u(0) = 10$, and $J(\gamma, 1) = \frac{(1-\gamma)^2}{10}$.

Taking beliefs as $b^{S_1S_2}_t = 0.25$ for all states, the optimal $\gamma^1_p = 1.355$, while $\gamma^2_p = 1.258$. Total welfare is given by 4.272.[41]

For comparison we consider the analogous context where the subconscious process can only bias γ_p for self (not for the teammate). In this case the optimal $\gamma_p = 1.163$. Total welfare in this case is given by 4.256.[42]

[40] This last inequality holds because $\frac{\partial \hat{b}^1_{t+1}}{\partial \gamma^1_s} \geq \frac{\partial \hat{b}^1_{t+1}}{\partial \gamma^2_s}$, i.e. the marginal benefit of distorting γ^1_s on own beliefs is always greater than those of distorting γ^2_s, combined with the fact that the mental cost function is identical for both γ^1_s and γ^2_s, and the material costs depend only on the ratio $\frac{\gamma^2_s}{\gamma^1_s}$.

[41] 0.759 utility from overconfident beliefs, 3.533 expected utility from material income, -0.019 dis-utility from mental costs.

[42] 0.733 utility from overconfident beliefs, 3.526 expected utility from material income, -0.003 dis-utility from mental

In the case of updating in a biased way about both teammates, the sub-conscious process becomes more biased about own performance, in order to gain from the benefits from overconfidence, whilst managing to lower the material costs.

B Direction of FAB for teammate 2

B.1 Theory

We reproduce the two updating equations for FAB, Equations 11 and 12 here.

$$\left[\hat{b}^2_{t+1}|s_t=1\right] = \frac{\gamma_p \Phi_{TT} b_t^{TT} + \Phi_{BT} b_t^{BT}}{\gamma_p \left[\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}\right] + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}}$$

$$\left[\hat{b}^2_{t+1}|s_t=0\right] = \frac{\gamma_n (1-\Phi_{TT}) b_t^{TT} + (1-\Phi_{BT}) b_t^{BT}}{\gamma_n \left[(1-\Phi_{TT}) b_t^{TT} + (1-\Phi_{TB}) b_t^{TB}\right] + (1-\Phi_{BT}) b_t^{BT} + (1-\Phi_{BB}) b_t^{BB}}$$

For arbitrary $\Phi_{S_1 S_2}$ we note that the posterior belief $[b^{2,FAB}_{t+1}|s_t = s]$ may in fact be greater than the Bayesian posterior.

Taking the derivative with respect to γ_p, γ_n:

$$\frac{d\left[\hat{b}^2_{t+1}|s_t=p\right]}{d\gamma_p} = \frac{\Phi_{TT} b_t^{TT} \cdot \Phi_{BB} b_t^{BB} - \Phi_{TB} b_t^{TB} \cdot \Phi_{BT} b_t^{BT}}{\left(\gamma_p \left[\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}\right] + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}\right)^2}$$

$$\frac{d\left[\hat{b}^2_{t+1}|s_t=n\right]}{d\gamma_n} = \frac{(1-\Phi_{TT}) b_t^{TT} \cdot (1-\Phi_{BB}) b_t^{BB} - (1-\Phi_{TB}) b_t^{TB} \cdot (1-\Phi_{BT}) b_t^{BT}}{\left(\gamma_n \left[(1-\Phi_{TT}) b_t^{TT} + (1-\Phi_{TB}) b_t^{TB}\right] + (1-\Phi_{BT}) b_t^{BT} + (1-\Phi_{BB}) b_t^{BB}\right)^2}$$

Note that in the experiment since there is symmetry $\Phi_{TT} = 1 - \Phi_{BB}$ etc. these conditions are equivalent.

This derivative will be negative, i.e. greater γ_s leads to a lower posterior about teammate 2 than Bayes' rule predicts if and only if $\Phi_{TT} b_t^{TT} \cdot \Phi_{BB} b_t^{BB} - \Phi_{TB} b_t^{TB} \cdot \Phi_{BT} b_t^{BT} \leq 0$.

Now we show that starting from independent priors, if individuals update according to this biased framework, this condition will hold whenever $\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT} \leq 0$. This is satisfied in our experiment $0.9 \cdot 0.1 - 0.5 \cdot 0.5 = -0.16 < 0$.

Denote prior beliefs by b_0^1, b_0^2. In the first round the performance of both teammates are independent, hence $b_0^{TT} = b_0^1 \cdot b_0^2$, $b_0^{TB} = b_0^1 \cdot (1 - b_0^2)$, and so on.

costs.

The expression of interest in the first round is thus:

$$\begin{aligned}\Phi_{TT}(b_0^1 \cdot b_0^2) \cdot \Phi_{BB}((1-b_0^1) \cdot (1-b_0^2)) - \Phi_{TB}(b_0^1 \cdot (1-b_0^2)) \cdot \Phi_{BT}((1-b_0^1) \cdot b_0^2) \\ = (b_0^1 \cdot b_0^2)((1-b_0^1) \cdot (1-b_0^2)) \cdot [\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT}] \end{aligned} \tag{24}$$

Thus, this expression will be negative, whenever $\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT} \leq 0$.
We now consider the next round of updating, after a positive signal is received.

$$[b_1^{TT}|s_t = p] = \frac{\gamma_p \Phi_{TT} \cdot b_0^{TT}}{\gamma_p[\Phi_{TT} \cdot b_0^{TT} + \Phi_{TB} \cdot b_0^{TB}] + \Phi_{BT} \cdot b_0^{BT} + \Phi_{BB} \cdot b_0^{BB}}$$

We note that the denominator of beliefs for all four states will be identical. Denote it by $\mathcal{D}_1 = \gamma_p[\Phi_{TT} \cdot b_0^{TT} + \Phi_{TB} \cdot b_0^{TB}] + \Phi_{BT} \cdot b_0^{BT} + \Phi_{BB} \cdot b_0^{BB}$. We now substitute this back into the initial expression of interest:

$$\frac{1}{\mathcal{D}_1}\left(\Phi_{TT}^2 \gamma_p b_0^{TT} \Phi_{BB}^2 b_0^{BB} - \Phi_{TB}^2 \gamma_p b_0^{TB} \cdot \Phi_{BT}^2 b_0^{BT}\right)$$

We now note that this is simply an iteration of Equation 24. As such it reduces to:

$$= \frac{\gamma_p}{\mathcal{D}_1}\left((b_0^1 \cdot b_0^2)((1-b_0^1) \cdot (1-b_0^2)) \cdot [(\Phi_{TT} \cdot \Phi_{BB})^2 - (\Phi_{TB} \cdot \Phi_{BT})^2]\right) \leq 0$$

We continue this inductive process once more:

$$[b_2^{TT}|s_t = p] = \frac{\gamma_p \Phi_{TT} \cdot b_1^{TT}}{\gamma_p[\Phi_{TT} \cdot b_1^{TT} + \Phi_{TB} \cdot b_1^{TB}] + \Phi_{BT} \cdot b_1^{BT} + \Phi_{BB} \cdot b_1^{BB}}$$

Where we denote $\mathcal{D}_2 = \gamma_p[\Phi_{TT} \cdot b_1^{TT} + \Phi_{TB} \cdot b_1^{TB}] + \Phi_{BT} \cdot b_1^{BT} + \Phi_{BB} \cdot b_1^{BB}$ and so hence:

$$\begin{aligned}[b_2^{TT}|s_t = p] &= \frac{\gamma_p \Phi_{TT} \cdot \frac{\gamma_p \Phi_{TT}}{\mathcal{D}_1} b_0^{TT}}{\mathcal{D}_2} \\ &= \frac{\left(\gamma_p \Phi_{TT}\right)^2 \cdot b_0^{TT}}{\mathcal{D}_1 \cdot \mathcal{D}_2}\end{aligned}$$

Thus we arrive at the third term:

$$= \frac{(\gamma_p)^2}{\mathcal{D}_2 \cdot \mathcal{D}_1}\left((b_0^1 \cdot b_0^2)((1-b_0^1) \cdot (1-b_0^2)) \cdot [(\Phi_{TT} \cdot \Phi_{BB})^3 - (\Phi_{TB} \cdot \Phi_{BT})^3]\right) \leq 0$$

Following this process, assume the k^{th} posterior is given by:

$$[b_k^{TT}|s_t = p] = \frac{\left(\gamma_p \Phi_{TT}\right)^k \cdot b_0^{TT}}{\mathcal{D}_1 \cdots \mathcal{D}_k}$$

Then the $k+1^{th}$ posterior:

$$[b_{k+1}^{TT}|s_t = p] = \frac{\gamma_p \Phi_{TT} \cdot b_k^{TT}}{\gamma_p[\Phi_{TT} \cdot b_k^{TT} + \Phi_{TB} \cdot b_k^{TB}] + \Phi_{BT} \cdot b_k^{BT} + \Phi_{BB} \cdot b_k^{BB}}$$

In particular, the $k+1^{th}$ term of this inductive process is:

$$= \frac{(\gamma_p)^k}{\mathcal{D}_1 \cdots \mathcal{D}_{k+1}} \left((b_0^1 \cdot b_0^2)((1-b_0^1) \cdot (1-b_0^2)) \cdot [(\Phi_{TT} \cdot \Phi_{BB})^{k+1} - (\Phi_{TB} \cdot \Phi_{BT})^{k+1}] \right) \leq 0$$

We note that given $\Phi^{TT} \cdot \Phi^{BB} = 0.09$ and $\Phi^{TB} \cdot \Phi^{BT} = 0.25$, this expression is strictly negative for all positive integers k.

B.2 Empirical Results

Without making any assumptions on the updating process, we can also simply examine the value of the expression: $\Phi_{TT} b_t^{TT} \cdot \Phi_{BB} b_t^{BB} - \Phi_{TB} b_t^{TB} \cdot \Phi_{BT} b_t^{BT}$, given actual beliefs in the experiment, and check whether it is less than or equal to 0. In fact in only 2% of cases is this expression positive.

C 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 C.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, i.e. our Main treatment, 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 Main 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.

The result that there is some limited evidence of a larger weight after positive signals is consistent with the results on asymmetric updating. Since subjects were also positively biased in updating about their teammate, this creates an overall moderating effect: the positive bias for both teammates works to cancel out, producing a more moderate weight report. A slight effect for positive signals is consistent with the slight over-weighting of positive signals for self relative to teammate 2, while for negative signals there was no significant difference in the structural framework. In the Control treatment the responsiveness to feedback was balanced across both teammate 1 and 2, and for both positive and negative feedback. This is consistent with the results in Table C.1. Finally, 7% of observations involved the submission of 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.

Table C.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

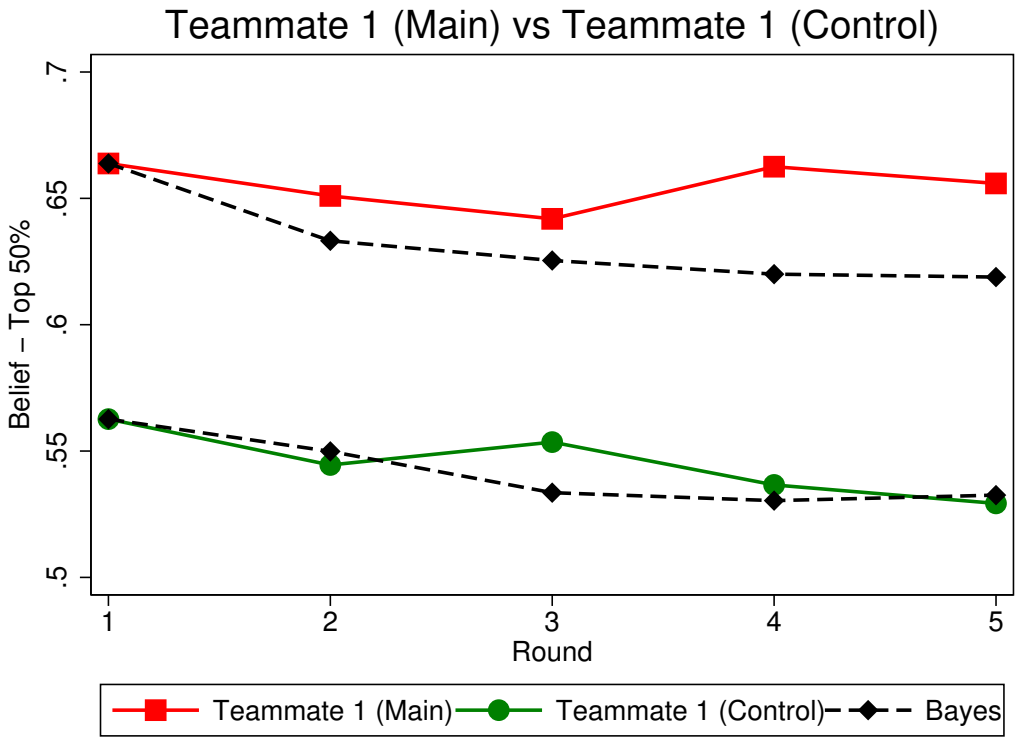
Analysis uses OLS regression. Difference is significant from 0 at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level.

D Examining Posterior Beliefs

Figures D.1 and D.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 D.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.

Figure D.1: Evolution of Beliefs: teammate 1

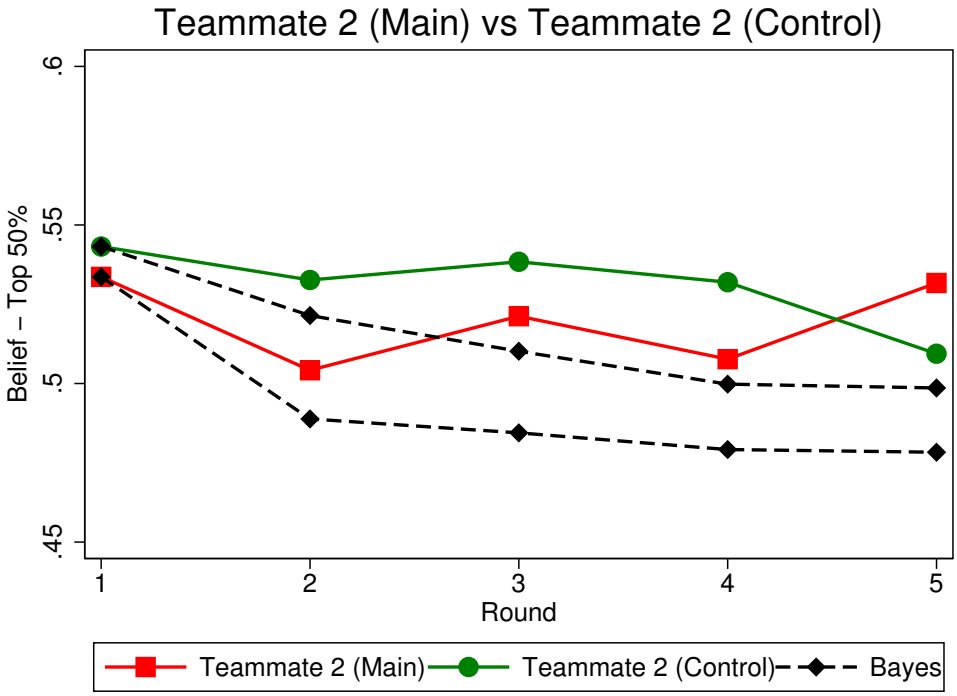


Evolution of beliefs about teammate 1 starting from prior beliefs with 4 round of feedback. Bayesian benchmark is calculated from subject's first prior, then evolves given actual signals observed. Standard error bars omitted for clarify (error bars are always overlapping with bayesian predictions).

Figure D.4 presents Epanechnikov kernal-weighted local polynomial smoothing plots regarding the relationship between priors and posteriors. The sample is identical to that of Tables 2 and 3. The Bayesian estimates are presented in black dashed lines, 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

Figure D.2: Evolution of Beliefs: teammate 2

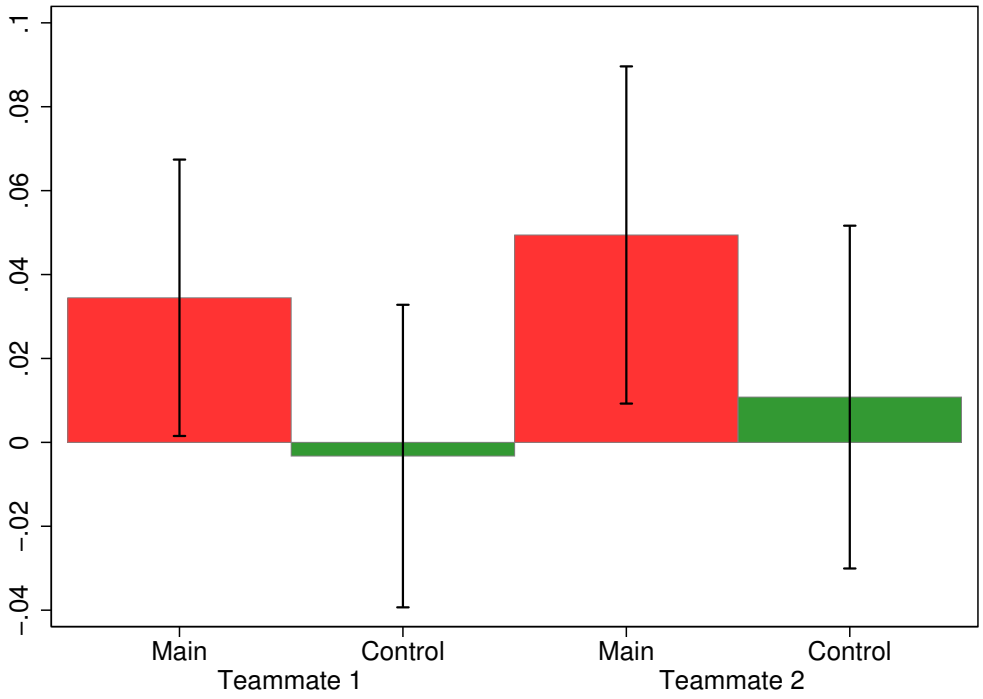


Evolution of beliefs about teammate 2 starting from prior beliefs with 4 round of feedback. Bayesian benchmark is calculated from subject's first prior, then evolves given actual signals observed. Standard error bars omitted for clarify (error bars are always overlapping with bayesian predictions).

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 are 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.

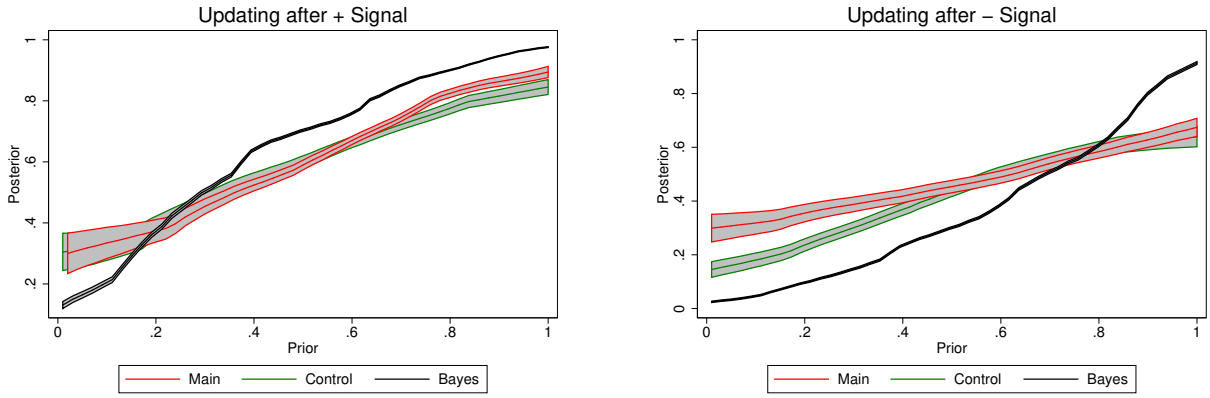
Figure D.3: Raw Deviation of Posterior Beliefs from Bayesian Benchmark



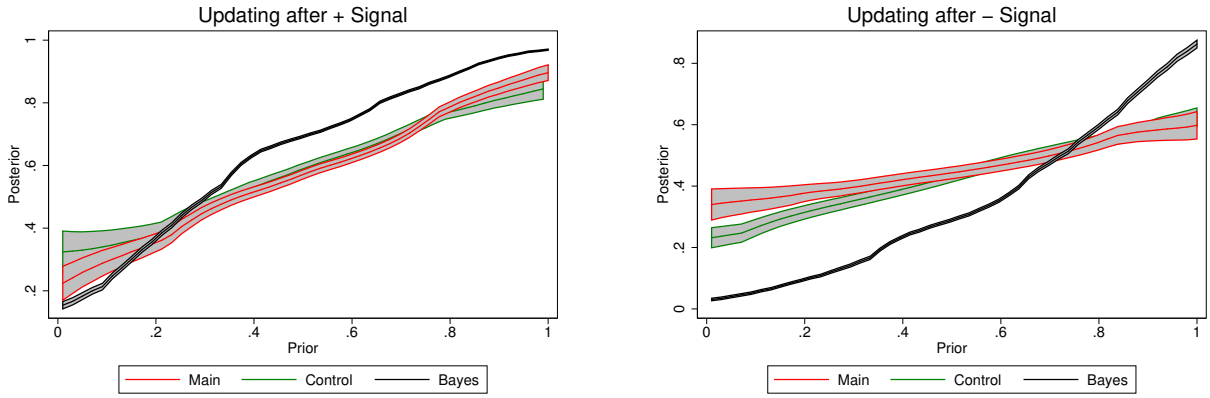
Plot of the difference between Posterior beliefs and Bayesian beliefs after 4 rounds of feedback. Bayesian beliefs are calculated using subject priors before any feedback.

Figure D.4: Relationship Between Priors and Posteriors

Teammate 1



Teammate 2



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

E WTP to Switch Teammates

In Wave 2 we provided subjects with the opportunity to be randomly re-matched 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.[43] After submitting their bid, the computer randomly generated a price, $p \in [0, 1]$ using a continuous distribution. 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. Before receiving feedback, this decision depends entirely on the belief about teammate 2. If a subject believes their teammate is in the top half with probability less than 50% they should pay to change, otherwise they should not be willing to pay any positive amount.[44]

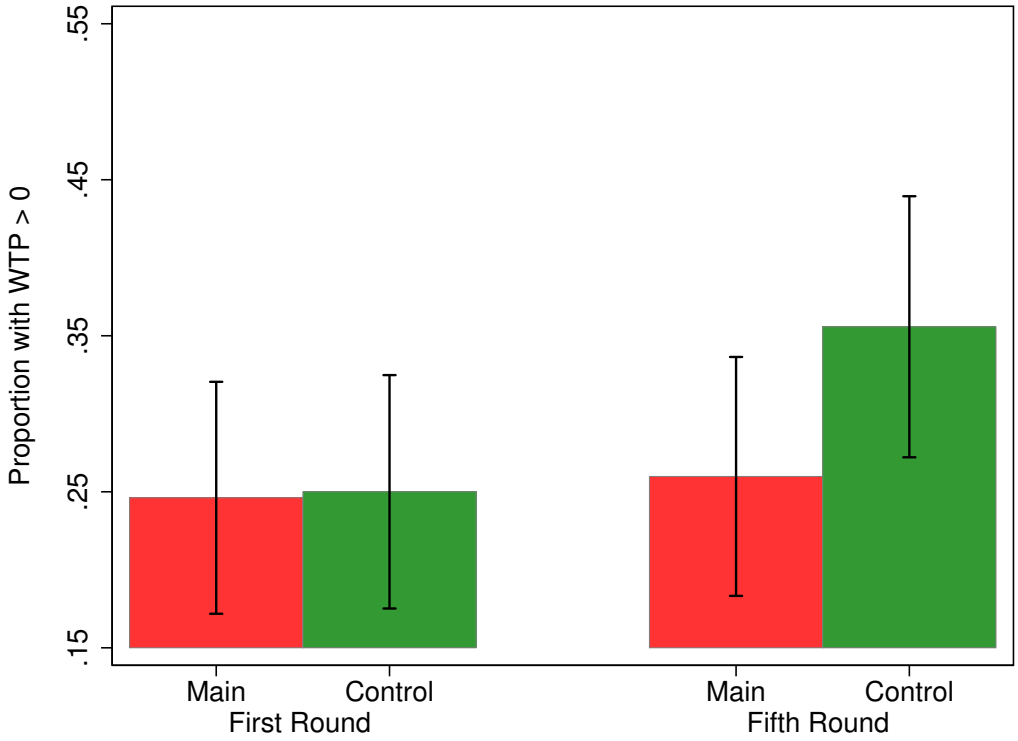
Since initial beliefs about teammate 2 are not statistically different across Main and Control treatments, we would predict that the number of subjects willing to pay a positive amount to change teammates will be the same across both groups. Figure E.1 confirms this is the case given prior beliefs in Main and Control (Round 1). This figure plots the theoretically optimal proportion of subjects which should opt to change teammates. While prior beliefs are such that there are no differences across Main and Control treatments, beliefs after 4 rounds of feedback (Round 5) are such that in fact that a higher proportion of individuals in Control should be willing to switch teammates. This is because in Control, subjects update in a symmetric way about their teammate, and end up with more moderate beliefs.[45] In Main, because of the asymmetry in updating about the teammate, there is no corresponding increase in the proportion that should switch teammates. As was shown in Figure 3, this is indeed the case for actual subject decisions.

[43] 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.

[44] One exception is if they believe with probability 1 that they themselves are in the top half, since they can choose a weight of $\omega = 1$ and mitigate any effect of a bad teammate. Note also that the *price* one is willing to pay is decreasing in beliefs about own performance. Higher performers are better able to hedge using their own performance, through choosing the optimal weight.

[45] In fact, since beliefs are initially slightly inflated about teammate 2, they end up with more pessimistic (but accurate) beliefs in Control.

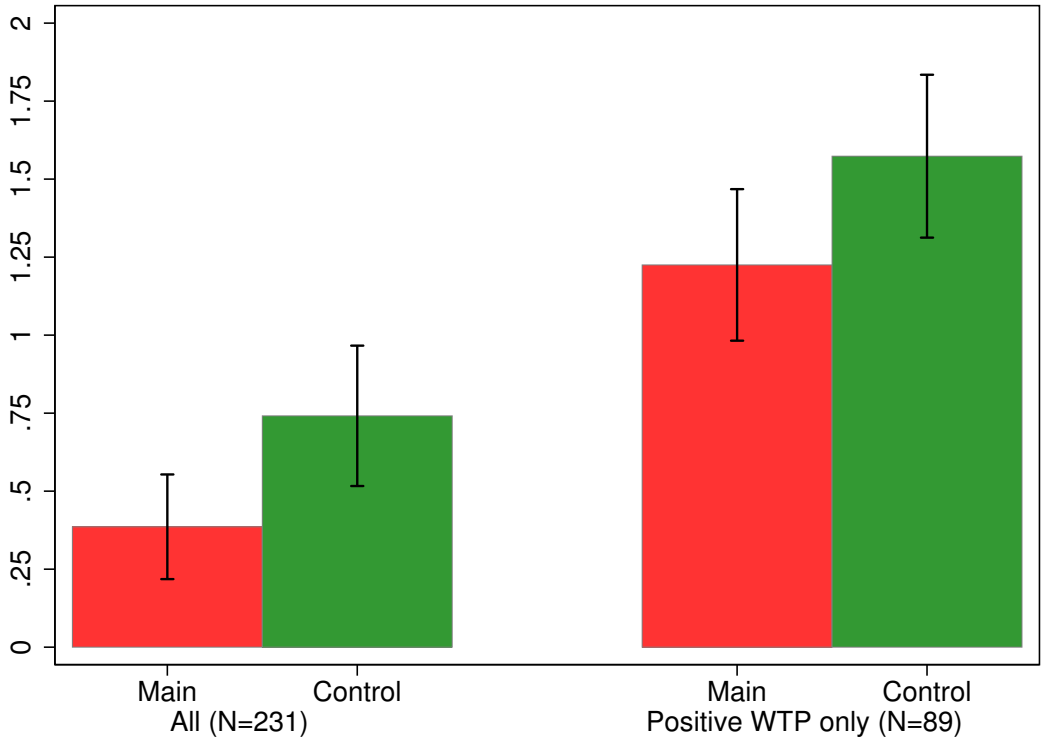
Figure E.1: (Calculated) Optimal Proportion Willing to Switch



Given subject beliefs, this shows the proportion of subjects that would (hypothetically) gain from switching teammates. 95% confidence intervals shown.

Figure E.2 presents the actual values of WTP submitted. The 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, $N = 89$. Thus while there is lower WTP among this restricted sample in Main treatment relative to Control, this can be accounted for by the more overconfident beliefs in Main, for which there is less material benefit to having a new teammate.

Figure E.2: Willingness to pay



WTP (in Euro) 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.