Biased belief updating and memory: The role of confidence

Julia Baumann[1,2,3] Ju Yeong Hong[1,2] Hedda Nielsen[1,2]

February 29, 2024

Abstract

We investigate the role of over- and underconfidence in belief updating and recall of feedback. While previous research finds mixed results for how positive and negative feedback impacts belief updating and recall, the overall focus has been on how (asymmetric) belief updating affects individuals' confidence. Instead, we ask how having high or low confidence impacts feedback processing. In an online experiment where we exogenously manipulate confidence levels, we find that underconfidence leads to less reaction to positive feedback compared to overconfidence in immediate belief updating. We do not find a treatment difference in the response to negative feedback. Further, there are no significant differences in feedback recall with respect to either under-/overconfidence or positive/negative feedback.

Keywords: Confidence, belief updating, memory

JEL Codes: C11, C90, D83

[1] Berlin School of Economics [2] Humboldt-Universität zu Berlin [3] Berlin Social Science Center (WZB). The authors thank Dorothea Kübler, Georg Weizsäcker, Anja Schöttner and seminar participants at HU Berlin and WZB for helpful conversations, the German Science Foundation for financial support via CRC TRR 190 (project number 280092119), the Einstein Foundation Berlin for financial support via the Einstein Visiting Fellowship of Bertil Tungodden and the Joachim Herz Stiftung. The design was approved by the institutional review board of Humboldt-Universität zu Berlin, Ethics Committee (Ethics Approval No. 2022-09) on 07.12.2022 and pre-registered at the AEA RCT Registry under AEARCTR-0012470.

1 Introduction

Beliefs about one's own ability – one's confidence level– are key to many economic decisions: from entering competition (Niederle and Vesterlund, 2007), to unraveling of job matching markets (Dargnies et al., 2019), or contractual choices (DellaVigna and Malmendier, 2006). To make these important choices, individuals have to form estimates of their true ability, which are oftentimes inaccurate. The size and direction of this inaccuracy can vary widely. Depending on the study and task, average priors sometimes indicate underconfidence (e.g. the verbal task in Ertac, 2011) and sometimes overconfidence (e.g. Zimmermann, 2020). If this initial heterogeneity in confidence shapes the way individuals process new information, it may also lead to important behavioral differences in economic decision-making.

Individuals frequently receive signals about their (relative) performance. Whether in the form of grades in school or performance evaluations at work, some signals are good news, others are simply bad news. While this should help in updating beliefs about (relative) ability –making individuals more or less confident–, receiving information about one's performance may not be sufficient to de-bias inaccurate beliefs, if biases are substantial or new information is not adequately incorporated.

Moreover, two people receiving the same feedback may react very differently to the same informational content. Consider, for example, a student with low self-confidence and a student with high self-confidence in their abilities, who work together on a team project. When their teacher tells them that their presentation does not measure up to the other teams' output, the low confidence student is much more inclined to agree with the criticism and take it seriously, while the high confidence student brushes it off as the teacher's bad mood or a matter of taste. This could imply that the low confidence student lowers their confidence even more while the other student's high confidence is not affected. Consequently, the two students' self-confidence levels are even more different than before the feedback. This example indicates a form of confirmation bias, predicting stronger updating of beliefs to expected signals than to unexpected signals. This type of bias deviates from the standard model of belief formation in economics where people incorporate new information following Bayes' Theorem (Benjamin, 2019). Put differently, the way individuals incorporate feedback may depend on their initial degree of confidence and whether or not the new information confirms, or dis-confirms, this initial image of oneself. To explore these considerations, we propose the role of confidence as one mediating factor that influences how individuals update their beliefs in response to feedback. Crucially, we compare belief updating to the Bayesian benchmark as much of the related literature.

When recalling the event at a later time, it may also be easier to remember the times you thought you did well, and did receive positive feedback, than the times when the information was unexpected and you doubted its veracity. The low confidence student from our example may naturally be more likely to recall their teacher's negative feedback because they reacted strongly to it when it was handed out. The high confidence student on the other hand may have forgotten all about it, given that they did not really pay attention to it in the first place. To explore this aspect of how we remember positive versus negative feedback, we also investigate the role of initial confidence in recall.

We can summarize our research question as: Does confidence causally impact feedback processing? We answer this question in two parts by examining (1) immediate belief updating after receiving feedback and (2) recall of feedback. To address this question, we run an online experiment with two main sessions: In Session 1, designed for investigating immediate belief updating, participants perform an ego-relevant, IQ-test style quiz and subsequently state their prior beliefs about their performance relative to individuals who performed the same task. Our treatment consists of randomly assigning either an easy or a hard version of the quiz to participants. This has been shown to induce relative over- and underconfidence (Moore and Healy, 2008) and has been used as a tool to exogenously manipulate confidence in other studies (Barron and Gravert, 2022; Dargnies et al., 2019). After the prior elicitation, we proceed to provide the participants with noisy feedback on their relative performance, after which we elicit their posterior beliefs. The noisy feedback is either positive or negative and always truthful. In the second session, after two weeks from Session 1, we ask participants to recall whether they received negative or positive feedback. We also observe patterns of belief updating in a non-ego-relevant setting in Session 2.

With our main sample ($N = 462$) we find that the *Hard* treatment that makes participants on average underconfident leads to less reaction to positive feedback compared to the *Easy* treatment, where participants are on average induced to be overconfident. This implies asymmetric, pessimistically biased belief updating with underconfidence, whereas reaction to positive and negative feedback is symmetric with overconfidence. We find no treatment difference in the reaction to negative feedback. To test whether the relationship between underconfidence and underreaction to positive feedback is simply an artifact of low priors, we task participants with a more neutral, non-ego-relevant belief updating task. We conclude that ego-relevance seems to be a necessary condition for the effect we find. Regarding memory of feedback, participants in our experiment were likely to correctly recall their feedback. They are not significantly more or less likely to recall their feedback by treatment group or type of feedback.

Thus, our results speak to the literature on belief updating where previous findings are mixed: Some studies find that individuals react differently to positive versus negative feedback (e.g. Möbius et al., 2022; Coutts, 2019; Eil and Rao, 2011; Ertac, 2011), while others do not find such an asymmetry (Barron, 2021; Coutts, 2019), or find asymmetry only after time has passed (Zimmermann, 2020).

Moreover, updating behavior that diverges from theoretical Bayesian benchmarks may have more persistent effects through the way feedback is recalled. Feedback recall has been shown to be important for understanding subsequent choices and beliefs about our own abilities (Zimmermann, 2020; Coffman et al., 2021; Huffman et al., 2022). For this purpose, we study if valenced feedback is recalled with a bias given prior confidence levels.

While confidence about one's performance has been found to matter for many different behaviors, evidence on the connection between confidence and updating remains sparse. Prior confidence about one's own (relative) performance is quite heterogeneous across studies (Ertac, 2011; Zimmermann, 2020). While other studies have studied the link from feedback processing to self-confidence (see in particular Möbius et al., 2022; Grossman and Owens, 2012), we examine the link from confidence to feedback processing with an exogenous manipulation to prior confidence.

Hence, we contribute to the literature by providing causal evidence between confidence in an ego-relevant task and belief updating, finding less reaction to positive feedback with underconfidence, leading to pessimistic updating, and symmetric updating with overconfidence. This demonstrates that heterogeneity in priors across studies can potentially explain some of the differing results on belief updating between studies. We further contribute to the somewhat newer literature on recall in economics, also adding more findings to the mixed evidence in this area.

The remainder of this paper is structured as follows: Section 2 goes into more detail on how previous literature relates to this study. In Section 3, we outline a theoretical framework for behavioral Bayesian updating in our context. Sections 4 and 5 present the experimental design and descriptive statistics on the experiment, respectively. Section 6 presents the detailed hypotheses, empirical strategy, and results for belief updating; Sections 7 and 8 follow the same structure for recall and the non-ego-relevant updating task.

2 Related Literature

Firstly, our study contributes to the investigation of heterogeneity in belief updating. An expanding number of studies in economics has been examining belief formation in response to feedback. A particular question has been whether individuals respond asymmetrically to "good news" compared to "bad news", thus deviating from the Bayesian updating prediction that implies symmetric belief updating regardless of the valence of the signals. The results from this literature are diverse. Some papers do not find asymmetric updating (Barron, 2021; in the short run, Zimmermann, 2020, Benjamin, 2019), whereas others find optimistic updating (Möbius et al., 2022; Eil and Rao, 2011), and yet other studies find pessimistic updating (Ertac, 2011; Coutts, 2019). While a recent study by Drobner (2022) proposes differences in expected uncertainty resolution as (partial) explanation for some of the heterogeneity between the economics and psychology literature, existing theories on drivers of differences in results fail to fully explain all observed differences in updating behavior (see Barron (2021) for an overview).

Our present study examines one possible factor in asymmetric belief updating that has not been studied systematically yet: the role of confidence. Some of the prior literature has focused on how biased belief updating affects confidence. This evidence suggests that optimistically biased updating produces overconfidence, thus providing a possible explanation for the persistence of overconfidence (Möbius et al., 2022; Grossman and Owens, 2012). The reverse causal relationship from confidence to biased belief updating has not been sufficiently studied. There is correlational evidence in some studies that underconfident priors produce less optimistically biased updating compared to overconfident priors (Ertac, 2011; Zimmermann, 2020; Huffman et al., 2022; Coffman et al., 2021).

By introducing a treatment that exogenously induces under- and overconfidence, we examine if there is a casual effect of confidence on belief updating that drives prior results from the literature. In this regard, our study is also closely related to so-called prior-biased updating, which implies stronger reactions to signals that confirm individuals' priors. Evidence for such confirmatory bias has been found in Charness and Dave (2017), while other studies did not find evidence of prior-biased inference (Eil and Rao, 2011; Möbius et al., 2022).

Secondly, our study makes a contribution to understanding the persistent effects of confidence and biased belief updating. More recently, the literature has shed more light on dynamic updating behavior, focusing on recall of positive and negative events (Zimmermann, 2020; Coffman et al., 2021; Chew et al., 2020). In doing so, they have examined not only the posterior belief of interest, but also the *recall* of feedback. In these studies, there is a follow-up experiment some time (e.g. 1 month) after the initial experiment where subjects are asked to recall the feedback they had previously received. This design allows for testing how feedback is processed in the longer run and the role that memory plays in this setting. Zimmermann (2020) finds that subjects updated neutrally as an immediate response to feedback, but remembered their posterior less when they had received negative feedback compared to when they had received positive feedback. However, this positive recall bias disappeared when subjects where provided with a stronger incentive for correct recall. They reason that motivated memory only wins out as long as the monetary costs of having this type of self-serving memory are low.

Similarly, a study by Chew et al. (2020) documents false memory in favor of positive events in the lab. With field data, Huffman et al. (2022) show that store managers at a retail chain recalled positive feedback more than negative feedback, leading to persistent overconfidence. On the other hand, Coffman et al. (2021) find that subjects are more likely to recall negative feedback than positive feedback (different signals) in their follow-up session, finding no evidence of motivated memory. They suggest that the difference between their result and other papers such as Zimmermann (2020) might arise due to the fact that accurate beliefs can help improve payoffs in their context.

In summary, this study is a novel contribution to a research field that often finds heterogeneous results. By proposing confidence as a moderating factor in feedback processing, we attempt to shed further light on some of the underlying reasons for different conclusions in these previous findings. Further, we expand the scope of our investigation by examining persistent effects of confidence in belief updating by studying feedback recall.

3 Theoretical framework of belief updating

To guide our understanding of how individuals process new information, we adopt a generalized model of belief updating. Following Möbius et al. (2022), we expand on the model developed by Grether (1980), here allowing for preference-biased updating using the framework of Benjamin (2019) and incorporating our confidence manipulation and feedback. There are two possible states, corresponding to being ranked in the top half of one's group (denoted as H) and being ranked in the bottom half of one's group (denoted as L). The agent receives one of two possible signals $S \in \{Pos, Neg\}$ that correspond to positive and negative feedback described in Section 4.1. Posterior beliefs are then formed as follows:

$$\frac{\pi(H|S)}{\pi(L|S)} = \left[\frac{p(S|H)}{p(S|L)}\right]^c \cdot \left[\frac{p(H)}{p(L)}\right]^d \tag{1}$$

where $\pi(\cdot)$ is the (possibly biased) belief, $\frac{\pi(H|S)}{\pi(L|S)}$ is the posterior odds ratio, $\frac{p(S|H)}{p(S|L)}$ gives the likelihood ratio and $\frac{p(H)}{p(L)}$ is the prior odds ratio. The parameters c and d signify how strongly the agent takes the signal (c) and their prior (d) into account. In the special case of $c = d = 1$, we obtain Bayesian updating. A parameter of $d < 1$ indicates base-rate neglect, while a parameter of $d > 1$ implies confirmatory bias. A value of $c < 1$ would imply underinference from the signal, whereas $c > 1$ would imply overinference from the signal relative to the Bayesian benchmark.

Updating in Equation 1 does not differentiate between “good news" and “bad news", i.e., between receiving a positive ($S = Pos$) or a negative ($S = Neg$) signal. Instead, our belief updating framework can be extended to take into account that agents may react differently to different types of signals. Adopting the preference-biased inference approach of (Benjamin, 2019), we can allow for agents to react more strongly to either positive or negative signals, depending on their prior beliefs:

$$\frac{\pi(H|S)}{\pi(L|S)} = \left[\frac{p(S|H)}{p(S|L)}\right]^{\{I(S=Pos)\cdot c_{Pos}+I(S=Neg)\cdot c_{Neg}\}} \cdot \left[\frac{p(H)}{p(L)}\right]^{d}, \tag{2}$$

where $I(S)$ is an indicator variable for the type of feedback and c_{Pos} and c_{Neg} indicate the degree to which an agent takes positive and negative signals, respectively, into account. All else remains as in Equation (1). If $c_{Pos} > c_{Neg}$, an agent portrays optimism, and if $c_{Neg} > c_{Pos}$ he or she instead exhibits pessimism.

Moreover, beyond general optimism and pessimism, in order to identify the causal impact of confidence on updating, we also separate the reactions to positive and negative feedback depending on confidence level, as instrumented by our treatment variable ($T = Hard$ for Hard, $T = Easy$ for Easy). For brevity, we redefine c as in the following expression

$$\begin{aligned} c := \{&I(S = Pos) \cdot c_{Pos,Easy} \cdot I(T = Easy) + I(S = Neg) \cdot c_{Neg,Easy} \cdot I(T = Easy)+ \\ &I(S = Pos) \cdot c_{Pos,Hard} \cdot I(T = Hard) + I(S = Neg) \cdot c_{Neg,Hard} \cdot I(T = Hard)\} \end{aligned}$$

Table 1 summarizes interpretations of the updating parameters based on potential deviations from Bayesian updating. Estimating and comparing our parameters of interest, $c_{Pos,Easy}$, $c_{Pos,Hard}$, $c_{Neg,Hard}$, and $c_{Neg,Easy}$, allows us to conduct a test of asymmetric belief updating and further check whether this asymmetry is driven by variations in confidence.

Table 1: Parameter Interpretations

Belief Updating Form	Parameter Values
Bayesian Updating	$\forall c = d = 1$
Base-Rate Neglect	$d < 1$
Confirmation Bias	$d > 1$
Over-inference	$c_{S,T} > 1$ for $\forall S, T$
Conservatism	$c_{S,T} < 1$ for $\forall S, T$
Optimism	$c_{Pos,T} > c_{Neg,T}$ for $\forall T$
Pessimism	$c_{Neg,T} > c_{Pos,T}$ for $\forall T$
Prior-congruency	$c_{Pos,Easy} > c_{Pos,Hard}$ and $c_{Neg,Hard} > c_{Neg,Easy}$
Prior-incongruency	$c_{Pos,Hard} > c_{Pos,Easy}$ and $c_{Neg,Easy} > c_{Neg,Hard}$

4 Experimental Design

To quantify the causal impact of confidence on belief updating and recall, our experimental design must consist of three things: 1) an exogenous manipulation of confidence, 2) an (ego-relevant) belief-updating task, and 3) an exogenous variation in feedback. To achieve this, we employ a between-subjects experimental design consisting of two main sessions.[1]

In **Session 1**, participants perform an IQ-test type task; we elicit beliefs about relative performance (prior), provide them with feedback on performance, and elicit beliefs about relative performance once more (posterior). In **Session 2**, participants are asked to recall the first session and the feedback they received then.

Our main **treatment** exogenously varies individuals' prior confidence level by randomly assigning individuals to an easy versus a hard version of the IQ task, as proposed by Moore

[1] The project was pre-registered on the AEA RCT Registry under AEARCTR-0012470.

and Healy (2008).[2,3] Feedback is also exogenously varied by randomly assigning subjects to groups of participants and letting the performance of one randomly drawn group member, compared to the subject's, determine whether the subject receives either positive or negative **feedback**. While informative, the binary feedback contains only noisy information about true relative performance, allowing us to study potential asymmetries in belief updating across treatment groups.

Moreover, to study whether the updating patterns we find are an artifact of low and high priors, we study belief updating in a non-ego-relevant task with exogenously assigned priors in a within-subject design. We therefore also include a second, now **non-ego-relevant belief updating task** in Session 2. In this task, we assign each individual a prior identical to their intrinsic prior belief from the ego-relevant task.

Figure 1 below illustrates the flow of our experiment, and the following sections detail each stage:

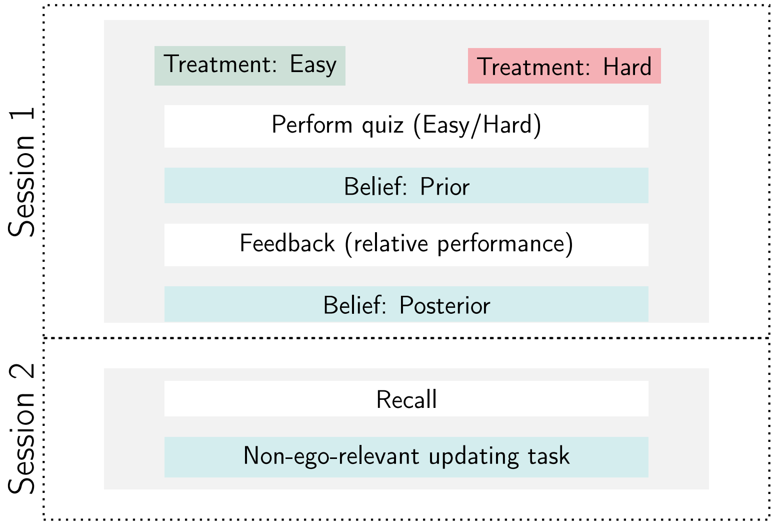

Figure 1: Experimental Design

[2] Generally, hard-easy manipulations such as this have been shown to induce relative under- and over-confidence with respect to one's belief about relative performance in a group, i.e., under- and over-placement (Bordalo et al., 2019).

[3] Full instructions along with the exact tasks used for each treatment can be found in the Appendix A.

4.1 Session 1

In Session 1, participants perform an ego-relevant, IQ-test style **quiz** consisting of a set of 10 Raven's Progressive matrices (Raven, 1965). For each matrix, participants are provided with eight possible answers, one of which is correct. Participants are given 10 minutes to complete the quiz and can freely decide how to allocate their time between tasks and revise any answer until the end of the 10-minute period. Participants are paid a piece-rate of GBP 0.10 for each correct answer.

After completing the quiz, or at the end of the allocated time, participants proceed to the **prior belief elicitation** stage. Participants are told that they are assigned to a group with five other randomly selected Prolific users who have taken the exact same quiz.[4]

We then elicit participants' (prior) beliefs about their performance relative to their assigned group members. We elicit beliefs in two parts: Firstly, our main belief of interest is the probability of scoring in the top half of one's group.[5] Secondly, we elicit the full belief distribution for each possible rank in the group (Rank 1 through Rank 6), in order to compute the Bayesian updating benchmark.[6] Here, we require that the beliefs for Rank 1, 2, and 3 sum up to the top half probability from the previous question. For all beliefs, participants are asked to report their estimates as an integer number between 0 and 100. We incentivize beliefs using the Binarized Scoring Rule (Hossain and Okui, 2013), following the recommended protocol in Danz et al. (2022) and randomly selecting either the first or the second belief to be payment relevant, yielding a bonus of GBP 0.50 if the section is chosen for payment.

Subsequently, participants are provided with noisy but informative binary **feedback** about their relative performance, consisting of a comparative signal of whether the participant has

[4] In order to allow for non-simultaneous completion of the study, we allocate participants to groups created from a reference group sample that was collected in earlier pilot studies. Importantly, the information is truthful: The groups consist only of participants who performed the exact same quiz, with the same set and order of questions.

[5] "What do you think is the probability that you ranked in the top half of your group, i.e., that you obtained Rank 1, 2, or 3, in your group of six?"

[6] "For each rank in your group of six, what do you think is the probability that you obtained this rank?"

performed better or worse than one randomly selected member of their allocated group. Concretely, one of the following two messages is shown: “You ranked better than your group member on the IQ quiz” or “You ranked worse than your group member on the IQ quiz”. Note that the feedback is always truthful. It is informative in the sense that being shown a positive message is more likely when the participant’s rank is better and vice versa. Henceforth, we define positive feedback as receiving a signal that one performed better than a group member and negative feedback as receiving a signal that one ranked worse than a group member.

Finally, we elicit participants’ **posterior beliefs** about their relative performance, asking the same two questions as in the first belief stage (prior) with the previous incentivization procedure and a bonus of the same size (GBP 0.50). For completing Session 1, participants are paid a show-up fee of GBP 1.50.

4.2 Session 2

After two weeks, participants who completed the first session are invited to Session 2.[7] For completing this session, participants earn a show-up fee of GBP 1.50.

In this session, we first ask participants to **recall** any aspect of the first session that they may still remember. They are then briefly reminded of the basic set-up of the study and asked to **recall** whether they received positive or negative feedback in the previous session[8]. Participants are paid GBP 0.25 if their answer is correct.

We then present a **belief updating** task in a **non-ego-relevant** context that is separate from the IQ quiz (“the bookbag-and-poker-chip” setting, Benjamin, 2019). Here, we display two bags, Bag A and Bag B, that contain red and blue balls, one of which will be chosen for each participant. If Bag A is chosen for a participant, the participant earns an additional bonus of GBP 0.30. Otherwise, they earn no extra bonus payment. A key element of this task is that we assign prior beliefs to the participants: Participants are informed of the

[7] Participants have 24 hours after the invitation to Session 2 to complete the study.

[8] “Did you rank better or worse than your group member on the IQ quiz?”

probability of Bag A being chosen, which is exactly equal to each participant's *prior belief about their relative IQ quiz performance* (i.e., the probability of scoring in the top half in one's group) as elicited in Session 1.[9] Assigning the exact same probability allows us to study whether updating in ego-relevant and non-ego-relevant contexts differs while holding the value of the prior constant.

We then provide the participants with noisy but informative feedback: We inform participants that a ball will be drawn from the bag chosen for each participant. While Bag A contains three blue balls and one red ball, Bag B contains one blue ball and three red balls.[10] If the ball drawn is blue, the signal is "good news" in the sense that it favors the state that comes with extra earnings (Bag A), while if the signal is the red ball, the new information is "bad news".

Finally, we ask participants two questions to elicit posterior beliefs about the probability of Bag A being chosen for a participant. Here, we utilize the strategy method[11] and elicit two beliefs, one for each possible signal, i.e., for good news and for bad news.[12] We incentivize this belief elicitation with the Binarized Scoring Rule and select one of the two questions for payment given the actual color of the ball drawn, rewarding a bonus of GBP 0.20.

[9] We adjust extreme beliefs as in the main data analysis: 0 to 1 and 100 to 99.

[10] Note that we keep the informativeness of the signal in this task similar to that of the feedback in IQ quiz belief updating by matching the signal factor of the blue ball with the average signal factor of positive feedback in our pilot data.

[11] We chose to implement the strategy method to ensure that we would observe updating behavior in response to the type of feedback each participant received in the main quiz, without having to resort to deception. For subjects who received positive feedback in Session 1, the first question relates to the "good news" signal; For subjects who received negative feedback in Session 2, the first question asks about the "bad news" signal.

[12] The two questions are as follows: "If a blue ball is chosen for you: The probability that Bag A was chosen for me is…" and "If a red ball is chosen for you: The probability that Bag A was chosen for me is…."

4.3 Additional Procedures

In addition to the procedures outlined above, we also included one further (un-incentivized) belief elicitation question in each belief stage — participants belief about their **absolute performance**[13] — as well as an un-incentivized **questionnaire** for mood, importance of IQ, previous experience with the task, risk-aversion, and demographic variables at the end of Session 1 and for mood and self-esteem at the end of Session 2. Table A1 in Appendix A provides an overview of the collected variables.

Moreover, in order to explore the potential impact of time on recall, we also invite our participants to complete a third session additional two weeks after Session 2. In this session, we repeat the recall section of Session 2, allowing us to study whether any results for recall are persistent over time.[14] In addition, we include a set of questions to measure Big 5 personality traits. Participants are paid a show-up fee of GBP 0.50, and the recall question is incentivized with a bonus of GBP 0.25 for a correct answer.

5 Sample & Descriptives

The experiment was conducted in November 2023 using experimental software oTree (Chen et al., 2016). Participants were recruited via Prolific and required to have a place of residence in the UK as well as at least a 95% approval rate on the platform. In addition, when recruiting, we ensure a balanced gender ratio. On average, it took subjects 15 minutes to finish Session 1 and 7 minutes to complete Session 2.[15]

[13] "Regardless of anyone else's performance in your group: How many tasks do you think you solved correctly?"

[14] This session was not included in our pre-analysis plan, and any results from it should thus only be considered exploratory.

[15] Session 3 lasted on average only 2 minutes.

5.1 Main Sample

As pre-registered, our **main sample** for our belief updating analyses consists of all participants who completed Session 1 and spent at least 30 seconds on the two pages where we elicited prior beliefs, which leaves us with a sample size of $N = 462$.[16] This includes participants both from our main data collection ($N = 393$) as well as from a pilot study in July 2023 ($N = 69$). Importantly, all procedures and instructions for the belief updating remain identical between these two sessions, and all main results remain qualitatively the same when excluding the pilot data from our sample.

Turning to recall and non-ego-relevant belief updating, our main sample consists of all subjects who returned for Session 2, leaving us with a sample of $N = 395$ for recall and $N = 340$ for non-ego-relevant belief updating.[17] As such, 84% of the participants returned to Session 2, and we find no selective attrition with respect to treatment group or feedback (see Table A3 for details).[18]

5.2 Prior confidence manipulation

To study the causal impact of confidence on feedback processing, our treatment must create exogenous variation in prior beliefs of our participants. Therefore, we start by examining the effect of our treatment manipulation on prior beliefs. Following the quiz, we elicit participants' beliefs about their probability of ranking in the top half of their group of six randomly assigned participants, i.e., of obtaining Rank 1, 2, or 3 ($b_i^{top} = Pr_i(\text{Rank}_i \in \{1, 2, 3\})$).

Figure 2 shows that the Hard-Easy manipulation was successful in separating relative performance beliefs between the two treatments. The mean prior belief in *Easy* was 60.97%, whereas the mean prior belief in *Hard* was significantly lower at 29.73% (two-sided t-test:

[16] As outlined in our pre-analysis plan, a sample size of $N = 450$ was chosen to allow us to detect a 0.2 difference in belief updating parameters between treatments at the 5% significance level with 80% power.

[17] The difference here comes from the fact that the July pilot did not include the non-ego-relevant task, yet all up until that point remained identical to the November data collection.

[18] For Session 3, we again invite all subjects who completed Session 2, yielding a total of $N = 298$ subjects.

$p < 0.001$). Not only are average beliefs significantly different from each other across the two treatments, they are also significantly different from 50%: The mean top half belief in *Easy* is significantly larger than 50% ($p < 0.001$), indicating *over*confidence; in the *Hard* treatment, the mean belief is significantly below 50% ($p < 0.001$), indicating *under*confidence.

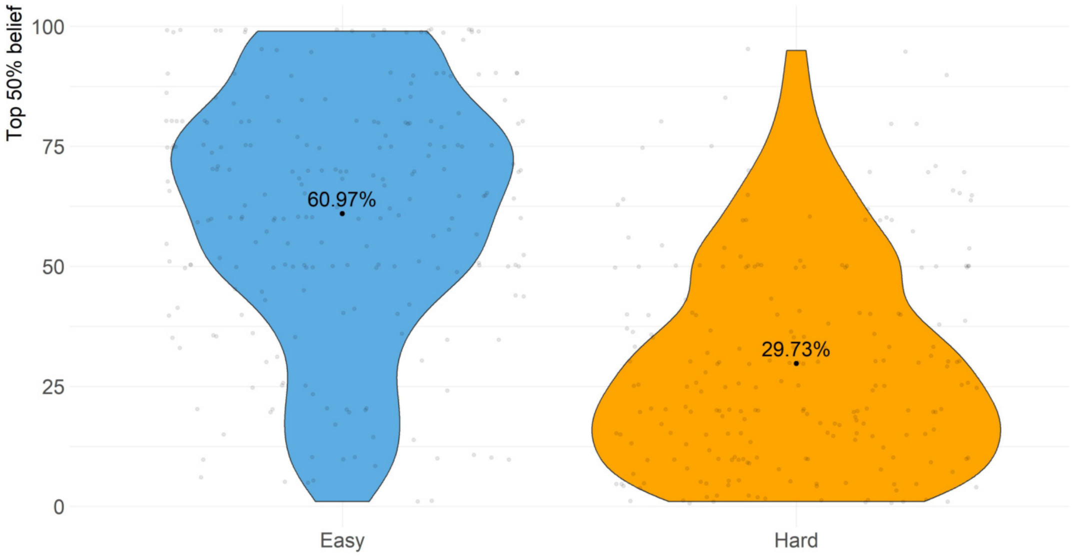

Figure 2: Prior belief for being in top 50% by treatment

We further simulate the actual probability of scoring in the top half for each score from 0 to 10. For each possible score, we repeatedly pick groups of six members and check each time whether a participant with this score would have scored in the top half of this group. The number of top half cases divided by the the total number of bootstrap samples ($G = 10,000$) yields the "actual top half probability" for each score and thus for each participant with that score. Based on this actual probability, we can compute the individual bias in relative confidence with: relative $\text{bias}_i = b_i^{top} - p_s^{\text{actual}}$. On average, this mistake is +9.1 in Easy and -19.6 in Hard.

5.3 Summary statistics

Table 2 shows summary statistics by treatment. In particular, the table demonstrates that the absolute quiz score as well as the belief discussed in the previous section are significantly different. Feedback is balanced in the sense that exactly 50% of participants received positive or negative feedback in each treatment.

Table 2: Summary statistics by treatment

Treatment		Easy			Hard		
Variable	N	Mean	SD	N	Mean	SD	Test
Actual quiz score (0-10)	234	7.2	2.2	228	2.3	1.7	F=699.352***
Belief prior top 50% (0-100)	234	61	26	228	30	22	F=198.54***
Belief prior absolute score (0-10)	234	6.7	2.1	228	3.2	1.8	F=374.139***
Feedback	234			228			X2=0
... Negative	117	50%		114	50%		
... Positive	117	50%		114	50%		
Mood (1-5)	234	3.8	0.86	228	3.4	0.95	F=18.725***
General risk attitude (0-10)	198	5.2	2.4	195	5	2.4	F=0.789
Importance of high IQ (1-5)	234			228			X2=4.464
... Very important	9	4%		13	6%		
... Important	26	11%		29	13%		
... Neither important nor unimportant	83	35%		94	41%		
... Not important	93	40%		76	33%		
... Not important at all	23	10%		16	7%		
General self-esteem (1-5)	194	3.5	0.9	200	3.6	0.8	F=0.942

Statistical significance markers: * p<0.05; ** p<0.01; *** p<0.001

Other covariates were balanced across treatments, with the exception of mood, indicating that mood was slightly worse in the Hard treatment (3.4 in Hard on a scale from 1 (bad) to 5 (good) vs. 3.8 in Easy).

6 Immediate belief updating

Building on our successful manipulation of prior beliefs, we can empirically test the relationship between confidence and immediate belief updating in response to positive and negative feedback. In this section, we first outline our hypothesis, followed by the empirical strategy we use to test it, and finally, we present the main findings.

6.1 Hypotheses on belief updating

Based on Section 3, and in particular Equation 2, we formulate the following hypothesis on immediate belief updating:

Hypothesis 1 *Differences in confidence lead to asymmetry in immediate belief updating to positive and negative feedback:*

1. *Individuals update their beliefs more in response to* ***positive*** *feedback in the* ***Easy*** *treatment than in the* ***Hard*** *treatment:* $c_{Pos,Easy} > c_{Pos,Hard}$
2. *Individuals update their beliefs more in response to* ***negative*** *feedback in the* ***Hard*** *treatment than in the* ***Easy*** *treatment:* $c_{Neg,Hard} > c_{Neg,Easy}$

At the most extreme, we would find over-inference from prior-congruent feedback (positive in Easy treatment and negative in Hard treatment) and an under-inference from prior-incongruent feedback (negative in Easy treatment and positive in Hard treatment) feedback; $c_{Pos,Easy} > 1 > c_{Pos,Hard}$ and $c_{Neg,Hard} > 1 > c_{Neg,Easy}$. However, as prior research shows a general tendency for conservatism ($c < 1$), we do not provide a strict hypothesis for over- and under-inference and instead focus on relative treatment differences.

6.2 Empirical strategy for belief updating

In order to test our hypothesis, we follow Möbius et al. (2022) and transform Equation 2 to be able to estimate belief updating parameters with a linear regression model. Firstly, we log-linearize the equation. Secondly, in accordance with our transformation of the belief updating parameters c in Section 3, we include interactions between the treatment and positive/negative signals in our regression model.

Allowing for individual deviations from Bayesian updating, we thus obtain the following empirical model:

$$
\begin{aligned}
\text{log-posterior-odds}_i = \; & \beta_{Pos} \cdot I(S_i = Pos) \cdot \text{log-likelihood Pos}_i \\
& + \beta_{Neg} \cdot I(S_i = Neg) \cdot \text{log-likelihood Neg}_i \\
& + \beta_{Pos \times Hard} \cdot I(S_i = Pos) \cdot \text{log-likelihood Pos}_i \cdot Hard_i \\
& + \beta_{Neg \times Hard} \cdot I(S_i = Neg) \cdot \text{log-likelihood Neg}_i \cdot Hard_i \\
& + \beta_{Prior} \cdot \text{log-prior-odds}_i + e_i,
\end{aligned} \tag{3}
$$

where the components are constructed for each participant –based on the actual signal S_i participant i receives and the two main belief variables (for top half beliefs (b_i^{top}) and full belief distribution (b_i^r for $r \in [1, 6]$) for prior and posterior beliefs)– as follows:

- **log-posterior-odds$_i$ (Dependent variable)**: The logarithm of participant i's posterior belief ratio based on their posterior belief for ranking in the top half of their group $b_{i,posterior}^{top}$, given their signal S_i: $\log\left(\frac{\pi_i(H|S_i)}{\pi_i(L|S_i)}\right) = \log\left(\frac{b_{i,posterior}^{top}}{1 - b_{i,posterior}^{top}}\right)$.
- **log-prior-odds$_i$**: The logarithm of participant i's prior belief ratio: $\log\left(\frac{p_i(H)}{p_i(L)}\right) = \log\left(\frac{b_{i,prior}^{top}}{1 - b_{i,prior}^{top}}\right)$.
- **log-likelihood Pos$_i$** and **log-likelihood Neg$_i$**: The log-likelihood ratios vary for each participant as the likelihood of receiving a positive or negative signal depends on the rank that the participant has actually obtained in their randomly assigned group. E.g., when a participant is in rank 1, the probability of receiving a positive signal is 1; a participant in rank 2 would receive a positive signal with probability 4/5 because they scored better than four of their five group members, and so on. Utilizing the full belief distribution priors (b_i^r for $r \in \{1, ..., 6\}$) we calculate the likelihood ratios as follows

(see Chadd et al., 2023, for the same procedure):

$$\log\left(\frac{p_i(Pos|H)}{p_i(Pos|L)}\right) = \log\left(\frac{b^1_{i,prior} \times (5/5) + b^2_{i,prior} \times (4/5) + b^3_{i,prior} \times (3/5)}{b^4_{i,prior} \times (2/5) + b^5_{i,prior} \times (1/5) + b^6_{i,prior} \times (0/5)} \cdot \frac{1 - b^{top}_{i,prior}}{b^{top}_{i,prior}}\right),$$

$$\log\left(\frac{p_i(Neg|H)}{p_i(Neg|L)}\right) = \log\left(\frac{b^1_{i,prior} \times (0/5) + b^2_{i,prior} \times (1/5) + b^3_{i,prior} \times (2/5)}{b^4_{i,prior} \times (3/5) + b^5_{i,prior} \times (4/5) + b^6_{i,prior} \times (5/5)} \cdot \frac{1 - b^{top}_{i,prior}}{b^{top}_{i,prior}}\right)$$

- $\boldsymbol{Hard_i}$: Treatment variable equal to one if a participant is assigned to the *Hard* treatment and zero otherwise.

- $\boldsymbol{e_i}$: Represents the individual-specific error term.

By estimating the coefficients β_k from Equation 3, we can test whether differences in confidence induced by different task difficulties lead to asymmetric belief updating in response to positive and negative feedback. For the main treatment effect, Hypothesis 1 therefore translates into $\beta_{Pos \times Hard} < 0$, which implies more belief updating after positive feedback in *Easy* than in *Hard* treatment, and $\beta_{Neg \times Hard} > 0$, implying more updating in response to negative feedback in *Hard* than in *Easy* treatment. Moreover, this theory-based regression framework allows us to test whether participants are Bayesian updaters and if they update asymmetrically to positive versus negative feedback overall.

6.3 Findings on belief updating

After providing the informative but noisy signal, we elicit posterior beliefs. In general, most participants update in the correct direction: Only 30 participants update in the wrong direction, i.e., downward relative to the prior after receiving positive feedback or vice versa for negative feedback. At the same time, a substantial share of the sample do not adjust their beliefs at all ($N = 116$).

To test our main hypothesis, we compare actual belief updating with the Bayesian benchmark. Figure 3 depicts the deviations from Bayesian updating by treatment and feedback type. The left panel (a) shows the Bayesian and actual updating patterns across feedback types for the mean prior beliefs in our two treatment groups, while the right panel (b) shows the mean deviation from the Bayesian benchmark for each feedback and treatment pair. In all groups, updating is significantly different from the Bayesian benchmark. In particular, participants exhibit "conservatism", i.e., they do not update enough in response to signals relative to the Bayesian benchmark.

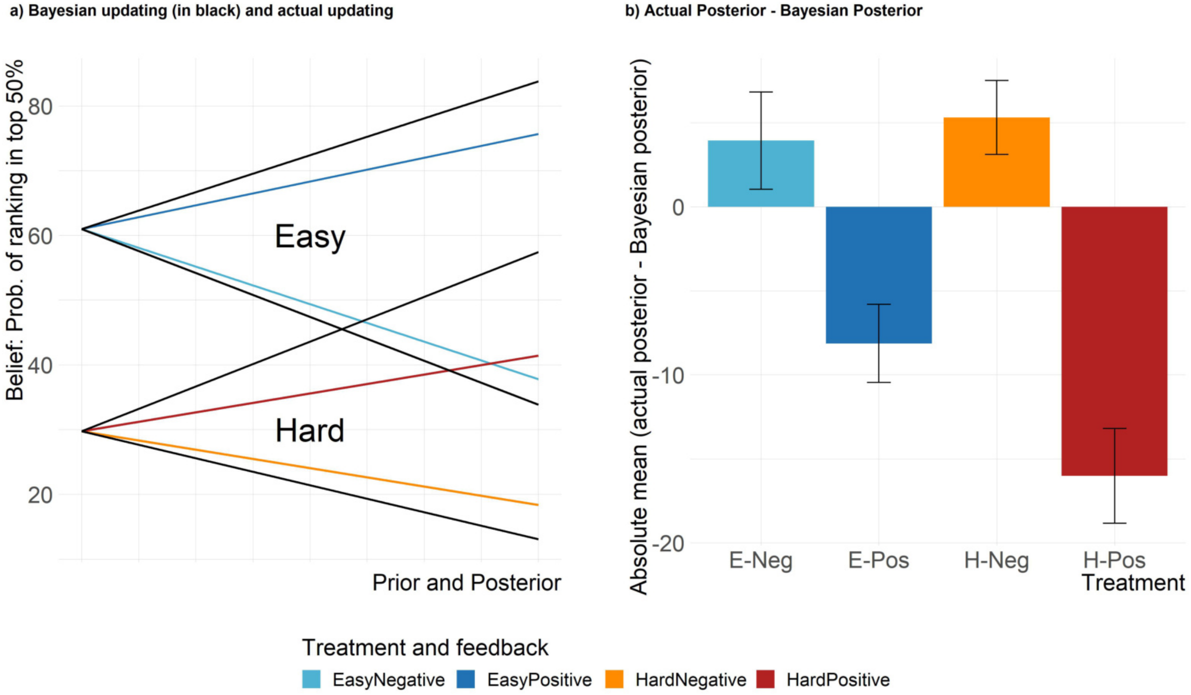

Figure 3: Actual vs. Bayesian updating by group

In terms of absolute belief adjustment[19], participants in the Easy treatment should on average revise their beliefs by 15.5 percentage points after positive feedback, but only change their beliefs by 7.4 points on average ($p < 0.001$, based on a two-sided t-test); for negative feedback, they should revise by 19.8 points, compared to 15.9 ($p = 0.01$) percentage points actual updating. In the Hard treatment, positive feedback implies 25.9 percentage points updating according to the Bayesian benchmark, but actual updating is a mere 9.9 points on average

[19] The absolute value of prior belief minus posterior belief.

($p < 0.001$); conversely, participants should have revised their beliefs by 14.9 points after negative feedback, but only do so by 7.4 points ($p < 0.001$).

Displayed conservatism is, however, less pronounced for negative feedback than for positive feedback. As can be seen in panel (b), the absolute difference between actual and Bayesian posteriors is smaller for negative feedback in both treatments. Pooling the treatments, the absolute gap between the Bayesian prediction and participants' posteriors is significantly larger for positive feedback: 14.8 versus 10.4 ($p < 0.001$, t-test).

Comparing our two treatments, for negative feedback, we find no significant difference in mean deviation from the Bayesian posterior (3.9 in Easy vs. 5.3 in Hard, $p = 0.453$, t-test unequal variance). For positive feedback, participants in Hard under-update significantly more than participants in Easy (-16.0 vs -8.1 resp., $p < 0.001$, t-test unequal variance).

The pattern from Figure 3 is confirmed by estimating the regression from Equation 3. On the one hand, Table 3 shows that the interaction between positive feedback (i.e., the log-likelihood ratio for positive feedback) and being in the Hard treatment is negative and statistically significant at (-0.291, $p = 0.005$). This is in line with Part 1 of Hypothesis 1 and implies that underconfident participants react less to positive feedback compared to overconfident participants. The weight participants put on positive feedback is estimated to be 0.592 in the Easy treatment, which is significantly below the Bayesian weight of 1 ($p < 0.001$). The weight on positive feedback is even lower in the Hard treatment (by 0.291), resulting in an estimated reaction to positive feedback of 0.301. More generally, Table 3 confirms that participants update conservatively, as all coefficients are significantly below 1 (the Bayesian benchmark). In Result 1 we summarize this first result:

Result 1 *When updating beliefs about being in the top half of their group, participants in Hard treatment infer significantly less from positive feedback compared to participants in Easy treatment. Both treatment groups update less than Bayesians, and this conservatism is significantly more pronounced for the Hard treatment group.*

	Dependent variable:
	logoddsPosterior
llr positive	0.592^{***}
	(0.077)
llr negative	0.667^{***}
	(0.065)
logoddsPrior	0.841^{***}
	(0.028)
llr positive x Hard	-0.291^{**}
	(0.104)
llr negative x Hard	0.066
	(0.101)
Observations	462
R^2	0.818
Adjusted R^2	0.816
Residual Std. Error	0.887 (df = 457)
F Statistic	411.322^{***} (df = 5; 457)
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 3: Belief updating regression - quiz, main results

On the other hand, the interaction effect between negative feedback and the dummy for being in the Hard treatment is insignificant and close to zero. We conclude that reaction to negative feedback does not depend on prior confidence levels in our sample. Thus, we find no evidence in favor of Part 2 of Hypothesis 1.

Result 2 *In response to negative feedback, participants update significantly less than Bayesians would do. We do not detect differences across treatment groups in belief updating to negative feedback.*

Figure 4 further depicts that there is no asymmetric updating in the Easy treatment but pessimistically biased updating in the Hard treatment. The figure plots the coefficients for the log-likelihood ratio for the two types of feedback as well as the log-prior-odds. While the estimated updating parameters on the log-likelihood ratio for positive and negative feedback overlap in the Easy treatment, the parameter estimation for positive feedback is substantially and significantly below the one for negative feedback in the Hard treatment ($p = 0.013$).

Result 3 *We find significantly asymmetric belief updating in the Hard treatment while updating in the Easy treatment shows no evidence of asymmetry.*

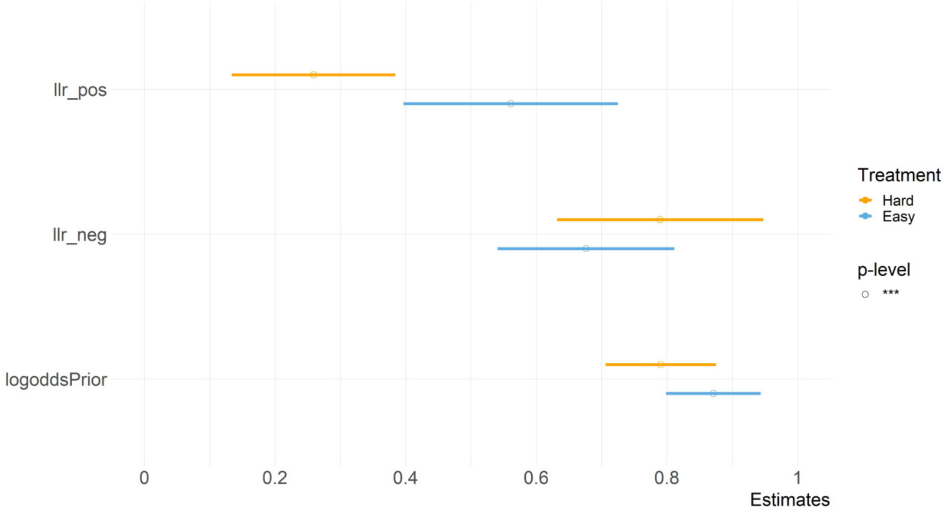

Figure 4: Updating about relative performance: Regression coefficients from Equation 3

For robustness, we repeat the main regression analysis with several pre-registered sample restrictions in Figure 5. None of the conclusions from Table 3 change with these restrictions. Further, we test the robustness of our results by adjusting extreme beliefs (0 and 100) using additional methods, as in Chadd et al. (2023). We present the results from these analyses in the Appendix.

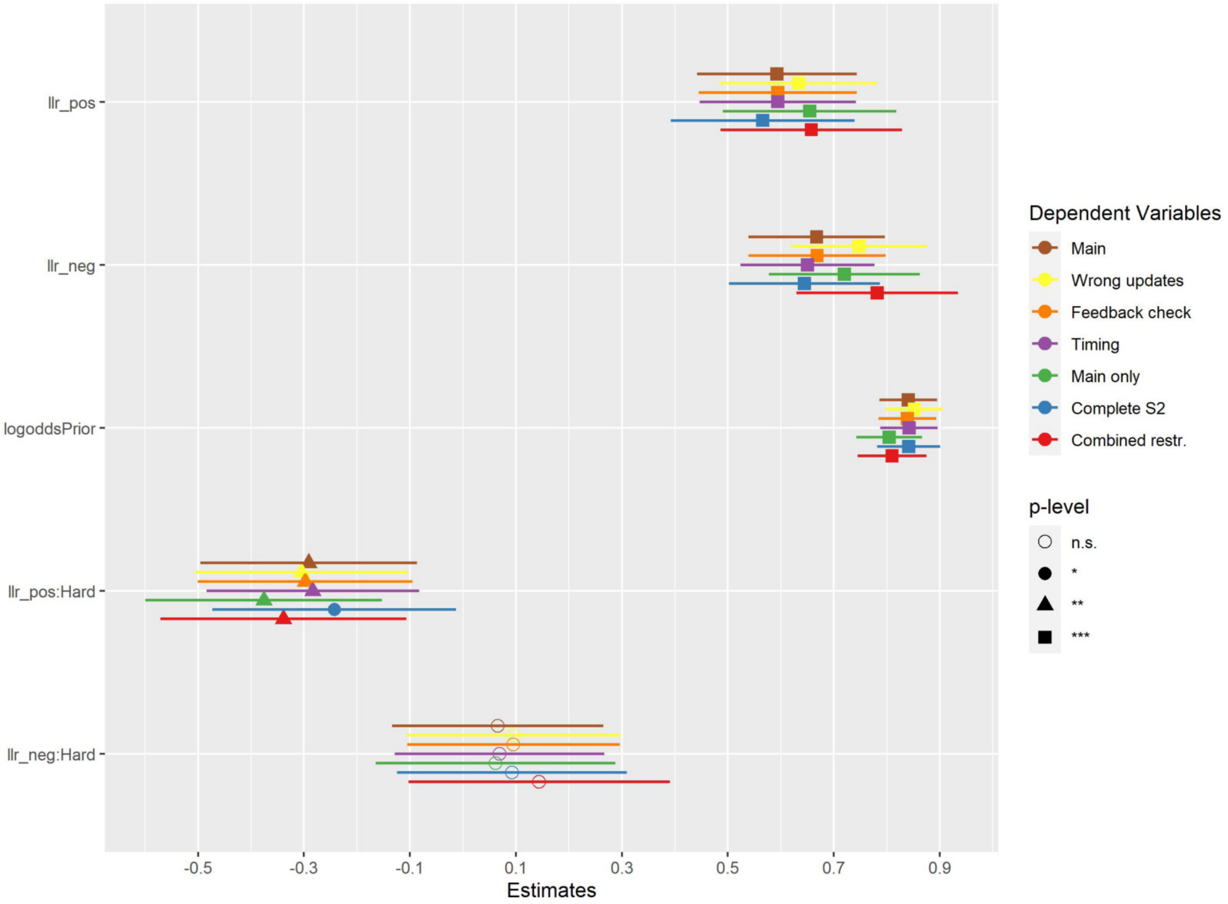

Figure 5: Regressions coefficients based on Equation 3 with varying sample restrictions
Sample restrictions are (as pre-registered): only participants who completed both Session 1 and Session 2 ($N = 394$); exclude participants who did no correctly answer an immediate attention check about the feedback they received ($N = 454$); exclude outliers in terms of completion time based on the interquartile range*1.5 ($N = 451$); exclude participant that update in the wrong direction (upwards with negative feedback or downwards with positive feedback; $N = 432$); exclude pilot data ($N = 393$); all restrictions combined ($N = 304$).

6.4 Heterogeneity

In this section we explore two main types of heterogeneity in treatment effects: Firstly, we show that deviations from the Bayesian benchmark differs depending on the relative, initial prior of the individual. Secondly, we test for heterogeneous treatment effects for belief updating and recall across a set of individual characteristics.

Deviations from the Bayesian benchmark by prior

Figure 6 below displays the deviation between the Bayesian prediction and the participants stated posteriors, split percentile of the prior belief distribution for each treatment.[20] As the left panel for positive feedback indicates, our main treatment difference seems to stem from the upper half of the belief distribution in each treatment. Here, individuals in the Easy treatment show significantly lower deviations from the Bayesian benchmark. As the right panel indicates, for negative feedback our result remains the same across the belief distribution.

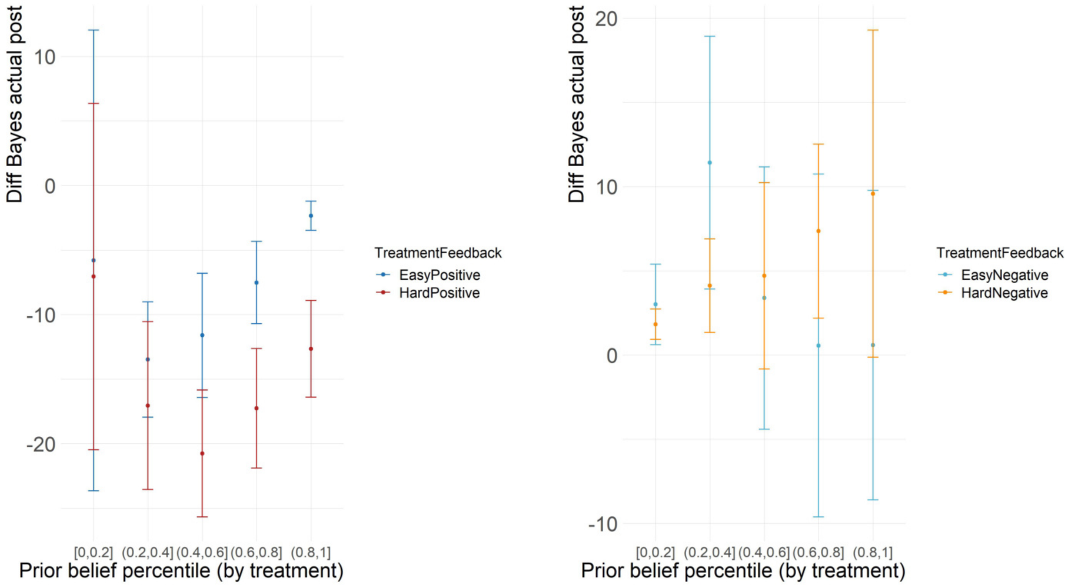

Figure 6: Updating by prior in the quiz

Belief percentile is defined by treatment group.

Subgroup analysis

We explore heterogeneity in belief updating and recall with respect to various covariates that we collect. We run sub-group regressions based on the regression models depicted in Equation 3, running separate regressions for two groups (mostly high/low levels) of each variable. Notably, there are no differences in our belief updating by gender, education, other

[20] Note: By splitting the sample by belief distribution we do not compare individuals who hold numerically same beliefs, but rather compare individuals with the relative same position in the confidence distribution.

demographic variables, or the quiz score (separated by the median score by treatment group). There is some suggestive evidence that the treatment effect may differ by self-esteem and risk preferences, but we are not powered to detect such effects. These results can be found in Figures A2, A3 and A4 in the Appendix.

7 Recall

In feedback processing, we are not only interested in immediate belief updating but also in more persistent effects of confidence regarding recall of feedback. To check this effect, we ask participants in Session 2 for the feedback that they were given in Session 1 and code their answer as a binary variable $\textit{correct recall}_i$ if they remember the feedback correctly.[21]

In Session 2, 394 participants out of the 462 participants returned. There is no selective attrition with respect to treatment, feedback, quiz performance, or prior beliefs. However, men, older participants, and risk-averse participants were slightly more likely to return to Session 2, factors that we control for in our regressions in this section (see Appendix, Table A3 for a detailed overview of differential attrition).

7.1 Hypotheses on recall

While the Bayesian updating framework does not incorporate imperfect recall, in reality, humans clearly do not always correctly recall information (e.g. Zimmermann, 2020). In the second session of our experiment, two weeks after the first Session, we therefore expect to see correct recall rates that are lower than 100%.

[21] The options were: "I ranked better than my group member", "I ranked worse than my group member", "I do not remember". Feedback is coded as correctly recalled when participants choose the option that corresponds to their actual feedback. "I do not remember" is always marked as incorrect.

Moreover, we would expect treatment differences in belief updating to carry over to differential recall rates of positive and negative feedback. In fact, even if we find no evidence of different belief updating between the Easy and Hard treatment groups, it is still possible that participants will exhibit differential biases in memory (similar to Zimmermann, 2020).

With this data, we investigate whether there is an asymmetry in recall, depending on the interplay of confidence level and the type of feedback. Namely, when we test the effect of confidence on feedback recall, we also consider whether or not the feedback is prior-congruent, i.e., positive feedback in Easy treatment (inducing high prior) and negative feedback in Hard treatment (inducing low prior). The hypotheses we test are as follows:

Hypothesis 2 *Prior-congruent feedback is more likely to be correctly recalled than prior-incongruent feedback – both within and between treatments:*

1. *Within* ***Easy*** *treatment, correct recall of positive feedback is more likely than that of negative feedback, while correct recall is more likely for negative than positive feedback in* ***Hard*** *treatment.*
2. *Across treatments, correct recall of positive feedback is more likely in* ***Easy*** *than* ***Hard*** *treatment, while recall of negative feedback is more likely to be correct in Hard than in* ***Easy*** *treatment.*

7.2 Empirical strategy for testing asymmetric recall

We test Hypotheses 2.1 and 2.2. with a two-sided t-test, comparing the means of the variable *correct recall*$_i$ by treatment and feedback. We also analyze the effect of confidence manipulation on correct recall using a linear probability model, controlling for the type of feedback and treatment as well as demographic and psychological measures we elicit, such as self-esteem, mood, and risk attitude.[22] We also control for whether participants correctly remembered at least a part of Session 1 (a binary variable coded as zero if a participant does

[22] See Table A2 for the full list of measures we elicit.

not remember anything, and one if correctly remembers at least a part of Session 1.). The model specification is as follows:

$$\text{correct recall}_i = \alpha + \alpha_{Pos} \cdot I(S_i = Pos) + \gamma \cdot Hard_i + \gamma_{Pos} \cdot Hard_i \cdot I(S_i = Pos) + b \cdot X_i + e_i, \quad (4)$$

where $Hard_i$ marks the treatment assignment, X_i is a vector of covariates, and e_i represents errors. We run this analysis both with and without the vector of covariates. By this estimation, we check the effect of feedback type on recall within treatments ($\alpha_{Pos} > 0$ and $\alpha_{Pos} + \gamma_{Pos} < 0$ for Hypothesis 2.1) and across treatments ($\gamma > 0$ and $\gamma + \gamma_{Pos} < 0$ for Hypothesis 2.2).

7.3 Findings on recall

It turns out participants recall the feedback given to them fairly precisely after two weeks, with an overall rate of correct recall reaching 88%. As depicted in Figure 7, we do not find any significant differences in the rate of correct recall by either treatment group or feedback. Aggregating the treatment groups, 87% of the participants who received positive feedback and 89% of those who received negative feedback remember the feedback type correctly ($p = 0.445$, t-test). While correct recall is slightly lower in the Hard treatment group with positive feedback than in the Easy treatment group with positive feedback (83% vs. 90%), this difference is not significant ($p = 0.121$, t-test unequal variance).

Although our preregistered hypotheses are focused on “correct” rather than “incorrect” memory across feedback types and treatment groups, when we compare wrong recall of feedback between treatments, we find that wrong memory of positive feedback—the case of remembering the feedback as being negative when one has actually received positive feedback—is more likely in the Hard (15%) than in the Easy (1%) treatment ($p = 0.0003$, t-test unequal variance). This could be taken as suggestive evidence that underconfident individuals are more likely to fabricate false memories of negative feedback.

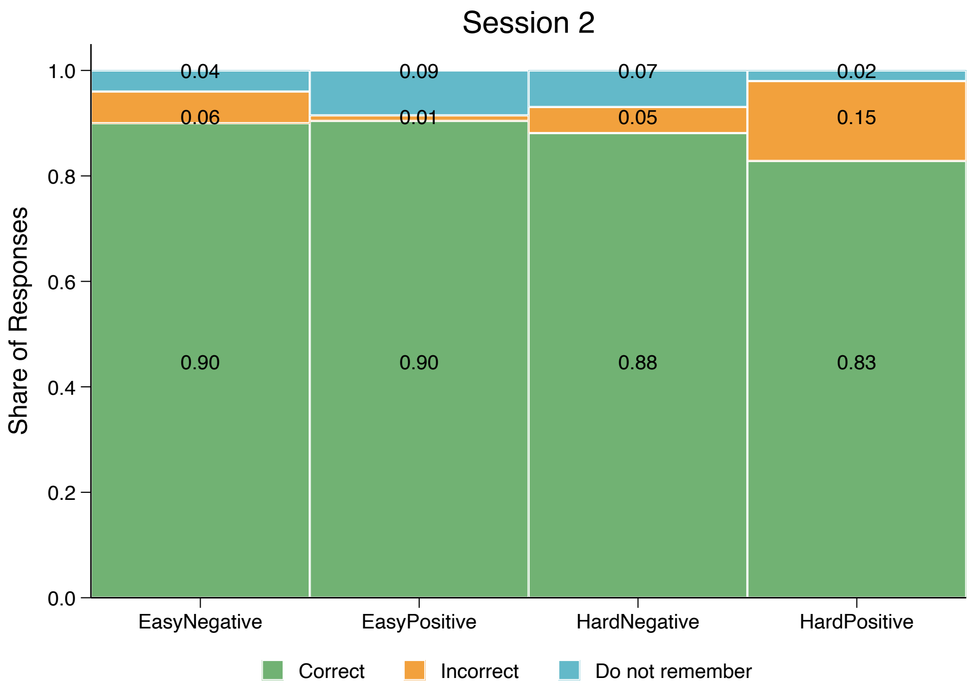

Figure 7: Average recall rates in Session 2 by treatment and feedback type

To study how recall evolves with time, we decided to run an additional, exploratory third session. In this session ($N = 298$), we measure the recall of feedback four weeks after the initial session and two weeks after Session 2. Overall we find that recall decreased slightly compared to Session 2, but still remains fairly high. The full shares are found in Figure A1. Again, we find no significant differences between treatment groups and feedback types. Subjects are generally consistent in their recall across sessions, with 87% reporting the same answer in both sessions.

Moreover, we find similar results on recall when we run regressions with and without demographic controls, as presented in Table 4. As such, the high level of recall across all groups does not seem to be driven by any particular group of participants. The baseline rate of correct recall is around 90% in Session 2 (estimation of constant in column (1) in Table 4) and 84% (estimation of constant in column (3) in Table 4) in Session 3.

Result 4 *We find no significant difference between recall of positive and negative feedback within either treatment. We also do not detect significant differences in correct recall across treatments, given the type of feedback.*

	Dependent variable:					
		CorrectRecall			CorrectRecall_3	
	Recall S2	Recall S2	Recall S2	Recall S3	Recall S3	Recall S3
	(1)	(2)	(3)	(4)	(5)	(6)
Hard	−0.019	−0.014	−0.032	0.050	0.067	0.068
	(0.046)	(0.048)	(0.054)	(0.057)	(0.059)	(0.059)
PosFeedback	0.004	0.022	0.029	0.050	0.066	0.054
	(0.047)	(0.049)	(0.054)	(0.057)	(0.059)	(0.060)
Hard:PosFeedback	−0.057	−0.064	−0.005	−0.131	−0.155	−0.151
	(0.066)	(0.068)	(0.076)	(0.081)	(0.083)	(0.084)
Constant	0.900***	0.775***	0.719**	0.842***	0.875***	0.846**
	(0.033)	(0.165)	(0.220)	(0.040)	(0.230)	(0.263)
Demographic controls	NO	YES	YES	NO	YES	YES
Personality controls	NO	NO	YES	NO	NO	YES
Observations	394	392	338	298	296	296
R^2	0.009	0.030	0.076	0.010	0.054	0.069
Adjusted R^2	0.001	−0.017	0.008	−0.0003	−0.004	−0.010
Residual Std. Error	0.327 (df = 390)	0.331 (df = 373)	0.339 (df = 314)	0.349 (df = 294)	0.350 (df = 278)	0.351 (df = 272)
F Statistic	1.116 (df = 3; 390)	0.639 (df = 18; 373)	1.122 (df = 23; 314)	0.970 (df = 3; 294)	0.932 (df = 17; 278)	0.873 (df = 23; 272)
Note:						*p<0.05; **p<0.01; ***p<0.001

Table 4: Recall regression

Subgroup analysis

Similar to our analysis of immediate belief updating, we run sub-group analysis for correct recall by estimating Equation 4 (without control variables). Results of these analyses can be found in Figure A5, A6 and A7. In general, there are no strong indications of heterogeneity in correct recall. However, participants who place a high importance on having a high IQ seem slightly more likely to remember positive feedback (Figure A5), yet this effect becomes insignificant once we correct for multiple hypotheses testing.

8 Non-ego-relevant belief updating

Using our main belief updating task about relative quiz performance, we find that underconfidence leads to less responsiveness to positive feedback. The remaining question is whether this finding results from ego-relevance of the task or it is a general phenomenon associated with low prior beliefs. To address this question, we look at the bag-and-ball belief updating task in Session 2, where we keep the level of prior belief of each individual the same as the reported prior about relative quiz performance in Session 1.

Given the same level of prior, we compare updating behaviors in ego-relevant versus non-ego-relevant contexts. We keep the structure of belief updating tasks alike, except for the ego-relevance, to make the two tasks comparable in other dimensions: Two states where one state (Bag A being chosen) is more desirable due to higher material benefit without any direct ego-enhancing component, two possible signals, one of which is good news and the other is bad news.

8.1 Hypothesis on belief updating in a non-ego-relevant task

As in Barron (2021), we expect deviations from the Bayesian benchmark to be smaller in the bag-and-ball belief updating task compared to belief updating about relative quiz performance:

Hypothesis 3 *Belief updating in the non-ego-relevant task deviates less from the Bayesian benchmark than belief updating in the ego-relevant (IQ) task.*

Overall, the non-ego-relevant task will allow us to investigate whether ego-relevance mediates the relationship between priors/confidence and updating. If ego-relevance matters for how priors impact belief updating, we will not see the same degree of asymmetric updating in the ego-relevant task and the non-ego-relevant task.

8.2 Findings on belief updating in non-ego-relevant task

We find that there is no difference in belief updating between treatment groups under our non-ego-relevant context. Table 5 and Figure 8 show this result. The two coefficients on positive and negative feedback are slightly smaller than their counterparts in Table 3, but similar. We find an even stronger tendency of base-rate neglect in the ball updating task than in the quiz updating task: The coefficient on the log-prior-odds is significantly smaller in the ball updating task compared to the quiz updating task. Overall, there is more noise in the updating process ($R^2 = 0.818$ in quiz updating vs. $R^2 = 0.459$ in ball updating). We see no asymmetric updating in the ball updating task.

	Dependent variable:
	logoddsPosterior_ball
llr positive	0.474***
	(0.125)
llr negative	0.589***
	(0.108)
logoddsPrior	0.487***
	(0.046)
llr positive x Hard	0.156
	(0.168)
llr negative x Hard	−0.211
	(0.169)
Observations	340
R^2	0.459
Adjusted R^2	0.451
Residual Std. Error	1.231 (df = 335)
F Statistic	56.921*** (df = 5; 335)
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 5: Belief updating regression - ball updating task

Thus, the evidence in favor of Hypothesis 3 is mixed: While updating is overall noisier and

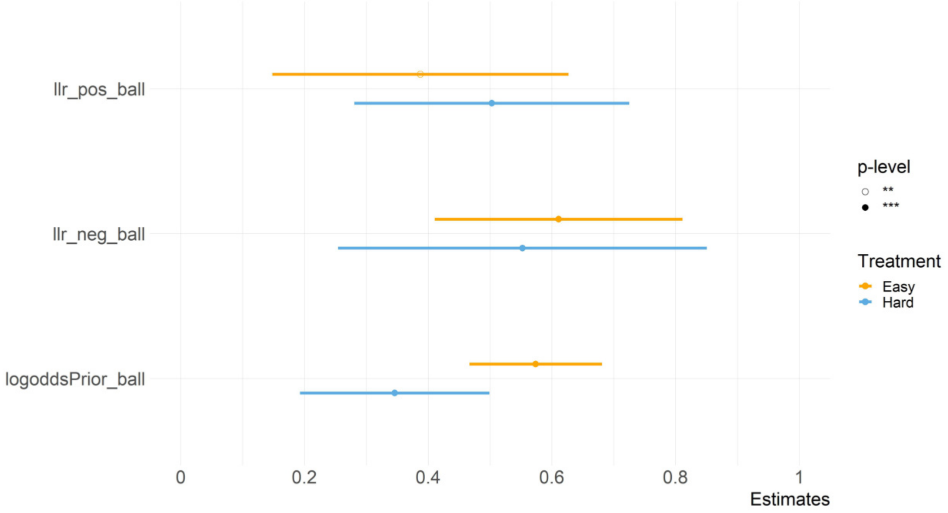

Figure 8: Updating in ball updating task: Regression coefficients from Equation 3

the estimated weight on the prior deviates more from Bayesian updating than in the ego-relevant task, there is no asymmetry in updating for either treatment group, which speaks in favor of more Bayesian-like updating.

Result 5 *We find no evidence of asymmetry in belief updating in the non-ego-relevant task. We detect a strong tendency of base rate neglect.*

Since we observe within-individual variations in updating for an ego-relevant and a non-egorelevant task, we can correlate updating behavior for each individual. In Figure 9, we can observe that there is only a weak correlation between deviations from the Bayesian benchmark in the quiz and in the ball updating task (Pearson's correlation coefficient: 0.148 ($p = 0.006$)).

Given these findings, we interpret pessimistic belief updating about relative performance in our Hard treatment as a behavior driven by ego-relevance of the task and not a mechanical artifact of low prior beliefs.

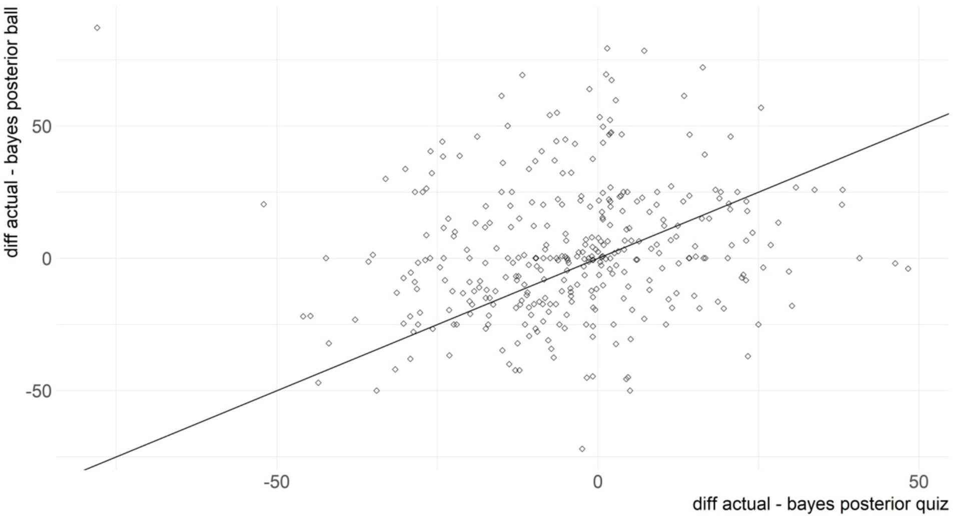

Figure 9: Within-individual deviations from Bayesian updating in the quiz vs the ball updating task

8.3 Heterogenous updating by prior

Figure 10 presents analogous results to Figure 6 for the non-ego-relevant ball updating task. While regression estimates did not show significant interaction effects in the ball updating task on average, this figure reveals that there are differences in the deviation from the Bayesian benchmark between treatment groups in some percentiles. Interestingly, the left panel shows an opposite pattern to what we find in the belief updating analysis for positive feedback: While deviations from the Bayesian benchmark are not significantly different for higher priors, they are different for lower priors. Moreover, lower priors in the Hard treatment tend to be associated with updating *over* the Bayesian benchmark, while lower priors in the Easy treatment are under the Bayesian benchmark; the complete opposite of what we find (and expected to find) in our ego-relevant main task.

For negative feedback, there is a general downward tendency in deviations from Bayes with increasing percentiles in the prior. Differences between treatment are less pronounced than for positive feedback, but seem to be increasing with the prior.

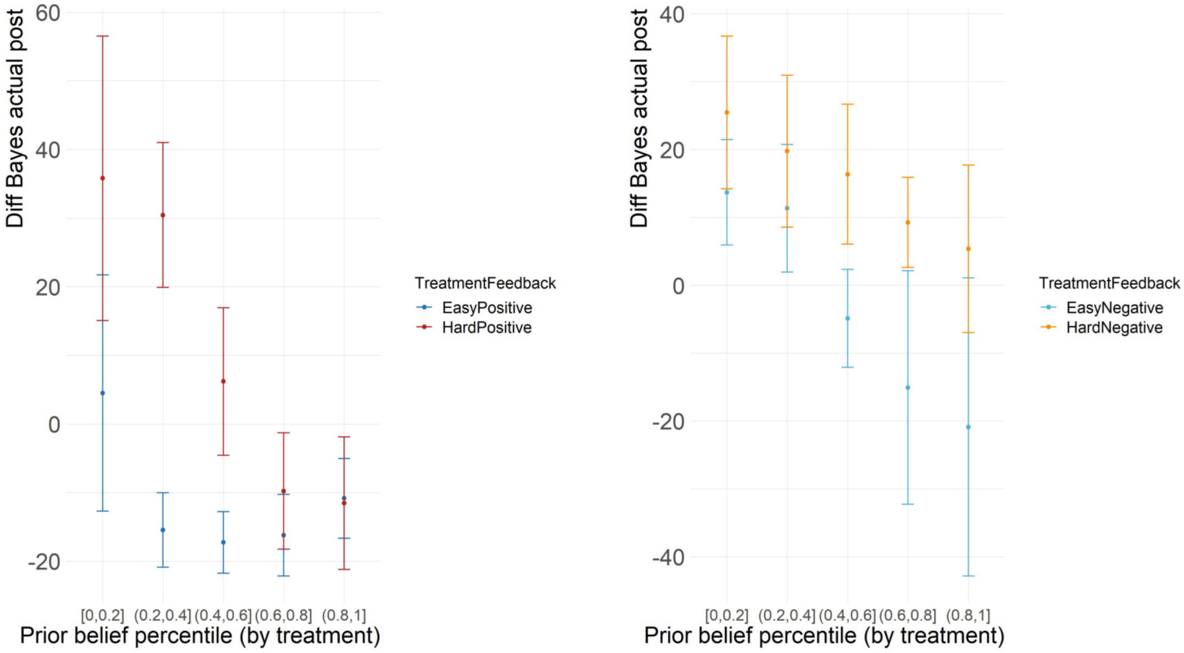

Figure 10: Updating by prior in the ball updating task
Belief percentile is defined by treatment group.

9 Discussion

All in all, our results highlight the importance of initial heterogeneity of confidence in the dynamics of belief formation. The degree to which an individual incorporates a new piece of information may fundamentally be different when their initial confidence is low or high. In this section we first place our results in the relevant literature, followed by discussing the implications and limitations of our results and design.

9.1 Connection to previous literature

As discussed in Section 2, the belief updating literature finds a general tendency for conservatism in belief updating. We also observe, similar to Möbius et al. (2022), conservatism in updating. Möbius et al. (2022) attribute this conservatism to a motivated bias, rather than cognitive bias, by showing that (i) more cognitively able participants update equally conservatively as less able participants, and that (ii) conservatism is less prominent, and the updating is closer to Bayesian benchmarks in non-ego-relevant setting where beliefs are about the performance of robots. Our data partially support this view: We find no differences in conservatism for individuals with above- versus below-median quiz performance or for individuals with high versus low levels of education.

On the asymmetry of belief updating, previous literature has come to mixed conclusions. With this study, We contribute evidence that some of the varying findings on asymmetry in belief updating may originate from different prior beliefs: On the one hand, priors in Zimmermann (2020) are on average overconfident, and he finds no asymmetry in immediate belief updating, a result we can essentially replicate in our Easy treatment. Moreover, in Ertac (2011), beliefs are often pessimistic and updating also pessimistic on average in an ego-relevant context, which mirrors the results from our Hard treatment. Chadd et al. (2023) finds women to be less confident than men on average, and that women tend to underweight good news when the signal has a comparative component – a result which is mirrored in our Hard treatment with positive feedback.

On the other hand, the results in Ertac (2011) also point to women being more pessimistic than men in settings where priors tend to be underconfident and Coffman et al. (2021) indicate that the gender-stereotype of the task may matter for updating behavior. We do not document any evidence of gender differences in either prior self-confidence or updating behavior. In the case of Coffman et al. (2021) this may however be due to different task choices, as Raven matrices are not strongly gender-stereotyped[23]. Similarly, Chadd et al. (2023) find that comparative signals –like ours– generate different updating for men and women than purely noisy signals. Together this confirms that while some aspects of belief updating can be explained by exogenously different levels of confidence, the process of belief updating itself may also be sensitive to specifics of the task and setting.

One key example of this is Drobner (2022), who documents optimistic updating when relative

[23] Previous studies have reported a a small male advantage only in a particular type of Raven matrix (Mackintosh and Bennett, 2005; Vigneau and Bors, 2008)

ability remains uncertain and updating patterns close to Bayesians' when the uncertainty about one's true rank is resolved immediately. As in the No Resolution treatment of Drobner (2022), our design also rules out uncertainty resolution, which leaves room for motivated updating. In contrast, however, we find symmetric updating under induced overconfidence, and pessimistic updating for participants with induced underconfidence. Importantly, compared to Drobner (2022), we manipulate the demand rather than the supply side of motivated updating. Posit for example that individuals who are overconfident to begin with, as in our Easy treatment, may not have any "demand" for ego-enhancing optimism, even if a "supply" side condition is relaxed due to no possibility of uncertainty resolution and hence, it is easier to manipulate the information processing. One caveat is that our data cannot provide clear evidence on this speculation since we do not observe optimism either when the initial confidence is induced to be low where the demand for optimistic updating might be larger. Other supply-side mechanisms than Drobner (2022)'s uncertainty resolution may therefore have played a role in restricting ego-boosting biases.

Another important example of this is ego-relevance. By including a non-ego-relevant task in our experiment and by keeping the initial priors fixed across the two tasks within each participant, we are able to conclude that individuals process similar types of signals (good/bad news, one signal) differently when the object beliefs are formed about is independent of the individual. Among studies of updating in non-ego-relevant settings, Möbius et al. (2022) achieves the non-ego-relevance by making the situation irrelevant to "self" but relevant to the performance of a robot. Instead, we isolate the ego-relevance by abstracting from the performance of any agents, human or mechanic. As such, we keep the good and bad dichotomous states of the world clearly relevant to self-benefit but irrelevant for ego-maintenance, ruling out any affective residue that may have developed for others who perform the same task as experimental subjects. As in the asymmetric treatments of Barron (2021), our two states differ only in terms of material benefit. The two possible types of signals exhibit a clear one-on-one relation to good/bad news since one signal involves higher benefit while the other does not. In contrast to Barron (2021), we do not provide multiple rounds of signals and also do not find updating close to Bayesian benchmarks. Instead, in our non-ego-relevant task we find substantial, symmetric deviations from Bayes'.

Moreover, beyond contributing to the growing literature on the particularities of asymmetries in belief updating, our paper also contributes evidence on the recall of feedback and on the role of attention in updating. Firstly, while we find no differences in recall by either feedback type or prior, Coffman et al. (2021) finds negatively biased recall and Zimmermann (2020) finds positively biased recall. Zimmermann (2020) further shows that motivated recall disappears when the incentive for recall is very high and, similarly, recalling feedback correctly can improve subsequent payoffs in Coffman et al. (2021). While recall is incentivized in our experiment, a reward of GBP 0.25 is arguably not large, and therefore unlikely to explain the lack of differences between groups and generally high rates of correctly recalled feedback. Instead, the nature of our feedback may have contributed to only finding directional evidence of asymmetry rather than substantial differences. With binary feedback and only one signal, it may be easier to remember (or to guess correctly) compared to more complex feedback types. In any case, the determinants of feedback recall are an area that future literature should expand on.

Secondly, in our setting, participants either exhibit confirmatory bias or react neutrally to signals. In the broader literature, both in economics (Bordalo et al., 2020) and in psychology (Filipowicz et al., 2018; Teigen and Keren, 2003) there exists the idea that surprising signals generate more attention. This contrast effect (Bordalo et al., 2020; Teigen and Keren, 2003) predicts more attention the further a new signal is away from what was expected. In this study, we observe less belief updating in response to surprising signals[24] and also significantly less correct recall of surprising feedback. This is not necessarily at odds with the contrast effect, as we do not directly measure attention. Rather, it may well be that participants in our sample paid a lot of attention to surprising signals, but then discarded them to some

[24] Surprise is defined as someone having a prior top half belief of above (below) 50% and then receiving negative (positive) feedback. Results on surprising feedback are available on request.

extent (especially with positive feedback in the Hard treatment), as showcased by less belief updating.

9.2 Implications and Limitations

Overall, we find that the degree to which positive or negative signals are incorporated when individuals form beliefs depends on their initial level of confidence, with two main implications: First, it emphasizes the gravity of external factors that shape our initial confidence levels. For example, being exposed to stereotypes that prescribe one's initial belief, e.g., gender stereotypes in different domains of knowledge as in Bordalo et al. (2019); Coffman (2014), can also shape the way we update beliefs by creating path dependence (Coffman et al., 2023).

Second, the main asymmetry we find is that individuals with average underconfidence react significantly less to positive feedback than individuals who are on average overconfident. If confidence carries over between tasks, then individuals who are once underconfident, may therefore retain their relative underconfidence, despite being provided with positive reassurances. Our findings on recall do not provide substantive evidence for differential rates of recall of feedback between confidence levels. However, it may still be that in a richer environment, where one forms beliefs about relative performance in multiple ego-relevant settings after one another, the underconfidence may carry over.

Beyond these direct implications, we also conclude that there is no evidence of ego-boosting behavior in our setting. Based on the vast literature on motivated beliefs (e.g. Zimmermann, 2020; Bénabou and Tirole, 2002), this may be a surprising result. However, as discussed in previous paragraphs, our study is generally in line with several other studies that do not necessarily find ego-enhancing beliefs. This is further evidence that while motivated reasoning (in the sense of ego-boosting) is undoubtedly an important driver of behavior in some contexts, this is far from a universal finding, which thus calls for a more nuanced and context-sensitive treatment of human belief formation.

Moreover, our study also has some natural limitations. Firstly, our design only allows us to estimate an intention-to-treat effect with our Hard-Easy manipulation, as our treatment by design does not provide a perfect and homogeneous movement of beliefs for every single participant. E.g. only 62% and 66% of participants are actually[25] over-/underconfident on the individual level in the Easy and the Hard treatment, respectively. Nonetheless, our treatment allows for the natural emergence of ("home-grown") priors, as would be realistic in most real-life settings where individuals would form different levels of self-confidence when faced with the same task or situation.

Secondly, our experiment naturally restricts attention to specific aspects of the relationship between confidence and feedback processing. Given that belief updating and feedback recall are clearly highly context-sensitive, further factors, such as the identity of the feedback provider, the nature of the task, and the role of emotions are likely to matter in this regard. Moreover, we focus on a particular type of relative confidence in our updating task, a belief about a performance ranking, whereas self-confidence could also be defined as a broader concept, leaving room for further explorations.

10 Conclusion

Our findings shed light on the causal effect of self-confidence on feedback processing. In a literature that has traditionally had to deal with mixed findings, we add results that have the potential to explain some of the different conclusions across studies. We are the first to document a causal relationship between underconfidence and the dismissal of positive feedback in belief updating. However, we also find that this effect is not necessarily long-lasting as individuals in our study do not display significantly different rates of feedback recall across prior confidence levels.

[25] This is calculated with the bootstrapped actual probability of ranking in the top half; see Section 5 for the exact definition of individual under- and overconfidence.

References

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

Barron, K. and Gravert, C. (2022). Confidence and career choices: an experiment*. *The Scandinavian Journal of Economics*, 124(1):35–68.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 69–186. Elsevier.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, Attention, and Choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.

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

Chadd, I., Osun, E. B., and Ozbay, E. Y. (2023). Effect of Feedback on Beliefs About Self-Ability.

Charness, G. and Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104:1–23.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chew, S. H., Huang, W., and Zhao, X. (2020). Motivated False Memory. *Journal of Political Economy*, 128(10):3913–3939.

Coffman, K., Collis, M. R., and Kulkarni, L. (2023). Stereotypes and belief updating. *Journal of the European Economic Association*, jvad063.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660.

Coffman, K. B., Araya, P. U., and Zafar, B. (2021). A (dynamic) investigation of stereotypes, belief-updating, and behavior.

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

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief Elicitation and Behavioral Incentive Compatibility. *American Economic Review*, 112(9):2851–2883.

Dargnies, M.-P., Hakimov, R., and Kübler, D. (2019). Self-Confidence and Unraveling in Matching Markets. *Management Science*, 65(12):5603–5618.

DellaVigna, S. and Malmendier, U. (2006). Paying Not to Go to the Gym. *American Economic Review*, 96(3):694–719.

Donnellan, M. B., Oswald, F. L., Baird, B. M., and Lucas, R. E. (2006). The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192.

Drobner, C. (2022). Motivated Beliefs and Anticipation of Uncertainty Resolution. *American Economic Review: Insights*, 4(1):89–105.

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

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

Filipowicz, A., Valadao, D., Anderson, B., and Danckert, J. (2018). Rejecting outliers: Surprising changes do not always improve belief updating. *Decision*, 5(3):165–176.

Grether, D. M. (1980). Bayes Rule as a Descriptive Model: The Representativeness Heuristic. *The Quarterly Journal of Economics*, 95(3):537–557.

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

Hossain, T. and Okui, R. (2013). The Binarized Scoring Rule. *The Review of Economic Studies*, 80(3):984–1001.

Huffman, D., Raymond, C., and Shvets, J. (2022). Persistent Overconfidence and Biased Memory: Evidence from Managers. *American Economic Review*, 112(10):3141–3175.

Mackintosh, N. J. and Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, 33(6):663–674.

Moore, D. A. and Healy, P. J. (2008). The Trouble with Overconfidence. *Psychological Review*, 115(2):502–517.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*, 68(11):7793–7817.

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

Peterson, R. A. and Sauber, M. (1983). A mood scale for survey research. In *Handbook of marketing scales*, pages 187–88. Association for Consumer Research.

Raven, J. C. (1965). Advanced Progressive Matrices. Sets I and II. London: HK Lewis & Co.

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package*, 61(52):18.

Teigen, K. H. and Keren, G. (2003). Surprises: low probabilities or high contrasts? *Cognition*, 87(2):55–71.

Vigneau, F. and Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, 36(6):702–710.

Zimmermann, F. (2020). The Dynamics of Motivated Beliefs. *American Economic Review*, 110(2):337–363.

A Appendix

A.1 Adjustment of extreme beliefs

As in previous literature, such as Grether (1980), Charness and Dave (2017) and Chadd et al. (2023), we adjust extreme beliefs (0 and 100) using the following protocol:

- For the "top half belief", we replace 0 with 1 percent and 100 with 99 percent.
- For the "full rank distribution", we apply the following procedure; Individual i's original belief for rank r is denoted by b_i^r, their adjusted belief for rank r is denoted by b_i^{r*}
    - Case 1: $b_i^{top} = 0$:

$$\begin{aligned} b_i^{r*} &= \frac{1}{3} \text{ if } r \in \{1, 2, 3\} \\ b_i^{r*} = b_i^r &- \frac{1}{3} \text{ if } r \in \{4, 5, 6\} \end{aligned}$$

    - Case 2: $b_i^{top} = 100$:

$$\begin{aligned} b_i^{r*} = b_i^r &- \frac{1}{3} \text{ if } r \in \{1, 2, 3\} \\ b_i^{r*} &= \frac{1}{3} \text{ if } r \in \{4, 5, 6\} \end{aligned}$$

    - Case 3: $b_i^{top} \in (0, 100)$:

- For $r \in \{1,2,3\}$:

$$\begin{aligned} b_i^{r*} &= \frac{1}{3} \text{ if } b_i^r = 0 \\ b_i^{r*} = b_i^r &- \frac{1}{6} \text{ if } b_i^r \neq 0 \text{ and } n_{0i}^{\text{top}} = 1 \\ b_i^{r*} = b_i^r &- \frac{2}{3} \text{ if } b_i^r \neq 0 \text{ and } n_{0i}^{\text{top}} = 2 \end{aligned}$$

- For $r \in \{4,5,6\}$:

$$\begin{aligned} b_i^{r*} &= \frac{1}{3} \text{ if } b_i^r = 0 \\ b_i^{r*} = b_i^r &- \frac{1}{6} \text{ if } b_i^r \neq 0 \text{ and } n_{0i}^{\text{bottom}} = 1 \\ b_i^{r*} = b_i^r &- \frac{2}{3} \text{ if } b_i^r \neq 0 \text{ and } n_{0i}^{\text{bottom}} = 2 \end{aligned}$$

Where n_{0i}^{bottom} (n_{0i}^{top}) describes the number of ranks with a zero probability for individual i in the bottom (top) half.

This adjustment prevents the undesirable generation of mechanical outliers during the log-linearization step in our main regression analyses outlined in Section 6. The method preserves consistency with the adjusted (or unadjusted) top half belief, such that the adjusted sum of rank 1,2 and 3 still sum up to b_i^{top}. In addition, the sum of all ranks still sums up to 100 for each individual. The method is inspired by Chadd et al. (2023); we also test if our main results are robust to the exact methods used in Chadd et al. (2023).

A.2 Overview of collected variables

Table A1: List of main variables

Variable	Definition	Scale	Elicitation timing
Score	Absolute performance that is measured by correctly solved numbers of Raven's Progressive matrices	0-10	Session 1
Prior belief - top half	Prior belief for the probability of ranking in the top half (Rank 1-3) of one's group	0-100	Session 1
Prior belief - full rank distribution	For each Rank 1-6, estimate of the probability for being in this rank.	0-100 for each rank. Must sum up to 100 and sum of Rank 1-3 must sum up to Prior belief top half.	Session 1
Prior belief - absolute performance	Prior estimate of correctly solved matrices	0-10	Session 1
Feedback	"You did better/worse than your [randomly chosen] group member."	positive/negative	Session 1
Posterior belief - top half	Posterior belief for the probability of ranking in the top half (Rank 1-3) of one's group	0-100	Session 1 , after feedback stage
Posterior belief - full rank distribution	For each Rank 1-6, estimate of the probability for being in this rank.	0-100 for each rank. Must sum up to 100 and sum of Rank 1-3 must sum up to Posterior belief top half.	Session 1
Posterior belief - absolute performance	Posterior estimate of correctly solved matrices	0-10	Session 1, after feedback stage
Recall Experiment	Open text field to describe what the participant recalls about Session 1	Open text field	Session 2
Recall Feedback	Participant is asked to indicate what feedback they where given in Session 1	Worse (1), Better (2), Do not remember	Session 2
Posterior Belief Bag A - good news	Probability that Bag A (the good state) was chosen given the signal is good news	0-100	Session 2
Posterior Belief Bag A - bad news	Probability that Bag A (the good state) was chosen given the signal is bad news	0-100	Session 2

Notes: We elicit participants' *Mood* both at Session 1 and Session 2 and *Self-esteem* at Session 2.

Table A2: List of survey variables, cont.

Variable	Definition	Scale	Elicitation timing
Gender	Gender one most identifies with	Male (0), Female (1), Non-binary or other (2), Rather not say (99)	Session 1
Age	Age	Years	Session 1
Education	The highest level of education one has completed	No formal qualifications (1), Secondary education (e.g., GED/GCSE) (2), High school diploma/A-levels (3), Technical/community college (4), Undergraduate degree (BA/BSc/other) (5), Graduate degree (MA/MSc/MPhil/other) (6), Rather not say (99))	Session 1
Ethnicity	Ethnicity indicated by a participant	African (1), . . . , Rather not say (99) [see table note for full scale]	Session 1
Income	Personal income per year, after tax, in GBP	Less than GBP 10,000 (1), GBP 10,000 - GBP 19,999 (2), GBP 20,000 - GBP 29,999 (3), GBP 30,000 - GBP 39,999 (4), GBP 40,000 - GBP 49,999 (5), More than GBP 50,000 (6), Rather not say (99)	Session 1
Ego-relevance	How important having a high IQ is for a participant	Very important (1), Important (2), Neither important nor unimportant (3), Not important (4), Not important at all (5)	Session 1
Experience with tasks	Whether one has previously solved the type of intelligence quiz in our study	Yes, many times (1), Yes, a few times (2), Yes, once (3), No (4), Do not remember (5)	Session 1
Risk aversion	An answer to the following question "How willing are you to take risks, in general?" adopted from a question in the German Socio-Economic Panel (SOEP).	Not at all willing to take risks (0) - Very willing to take risks (10)	Session 1
Mood	Average of four items measure of participant's current feeling, inverting reversed items (Peterson and Sauber, 1983)	Strongly disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly agree (5)	Both Session 1 & 2
Self-esteem	Average of 10 items measure of self-esteem, inverting reversed items (Rosenberg, 1965)	Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)	Session 2
Big 5	Average of 20 items measure of big five personality—openness, conscientiousness, extraversion, agreeableness, and neuroticism—inverting reversed items (Donnellan et al., 2006)	Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)	Session 3

Notes: Full scale for Ethnicity: African (1), Black / African American (2), Caribbean (3), East Asian (4), Latino / Hispanic (5), Middle Eastern (6), Mixed (7), Native American or Alaskan Native (8), South Asian (9), White / Caucasian (10), White / Sephardic Jew (11), Other (12), Rather not say (99). We elicit participants' *Mood* both at Session 1 and Session 2 and *Self-esteem* at Session 2.

A.3 Further results

A.3.1 Test for selective attrition: Session 1 vs. Session 2

Table A3: Attrition

complete		0			1		
Variable	N	Mean	SD	N	Mean	SD	Test
Treatment	68			394			X2=1.765
... Easy	40	59%		194	49%		
... Hard	28	41%		200	51%		
Feedback	68			394			X2=0.845
... Negative	30	44%		201	51%		
... Positive	38	56%		193	49%		
Actual score	68	5.3	3.1	394	4.7	3.1	F=2.337
Belief top 50%	68	46	28	394	45	29	F=0.065
Belief absolute score	68	5.4	2.5	394	4.9	2.7	F=1.713
Gender	68			394			X2=10.546*
... Male	25	37%		197	50%		
... Female	41	60%		193	49%		
... Other	2	3%		1	0%		
... Rather not say	0	0%		3	1%		
Age	68	34	10	394	42	14	F=19.988***
Education	68	4.4	1.2	392	4.3	1.3	F=0.218
Income	68	3.3	1.7	394	3.3	1.8	F=0.01
Mood	68	3.6	0.9	394	3.6	0.93	F=0
Importance of high IQ	68	3.4	0.84	394	3.3	0.97	F=1.277
General risk attitude	53	5.8	2.1	340	5	2.5	F=5.795*

Statistical significance markers: * $p<0.05$; ** $p<0.01$; *** $p<0.001$

A.3.2 Session 3 results

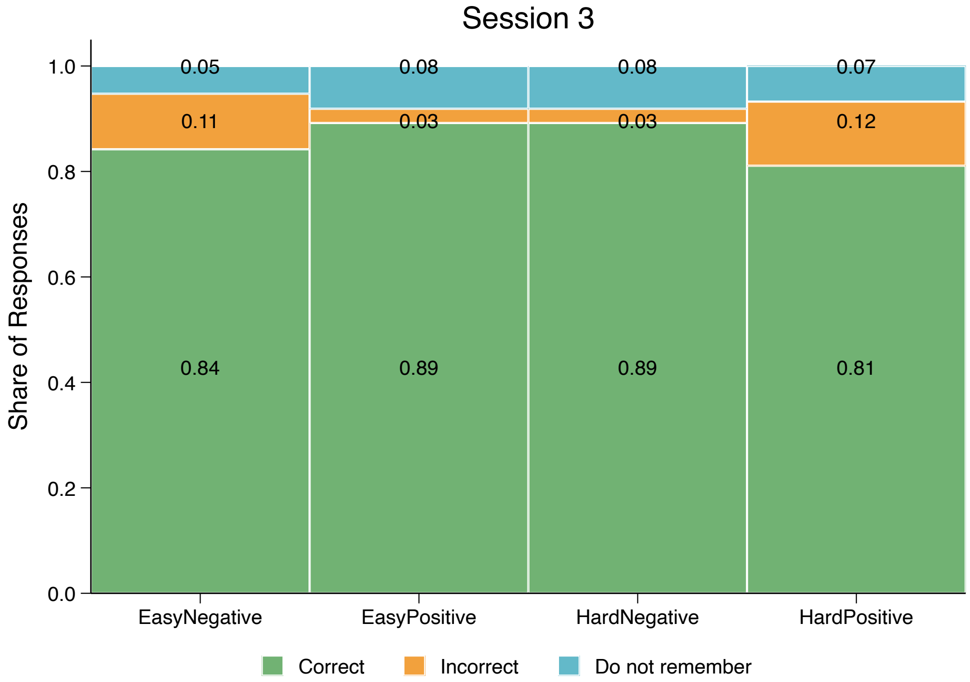

Figure A1: Average recall rates in Session 3 by treatment and feedback type

A.3.3 Subgroup analysis: Belief updating

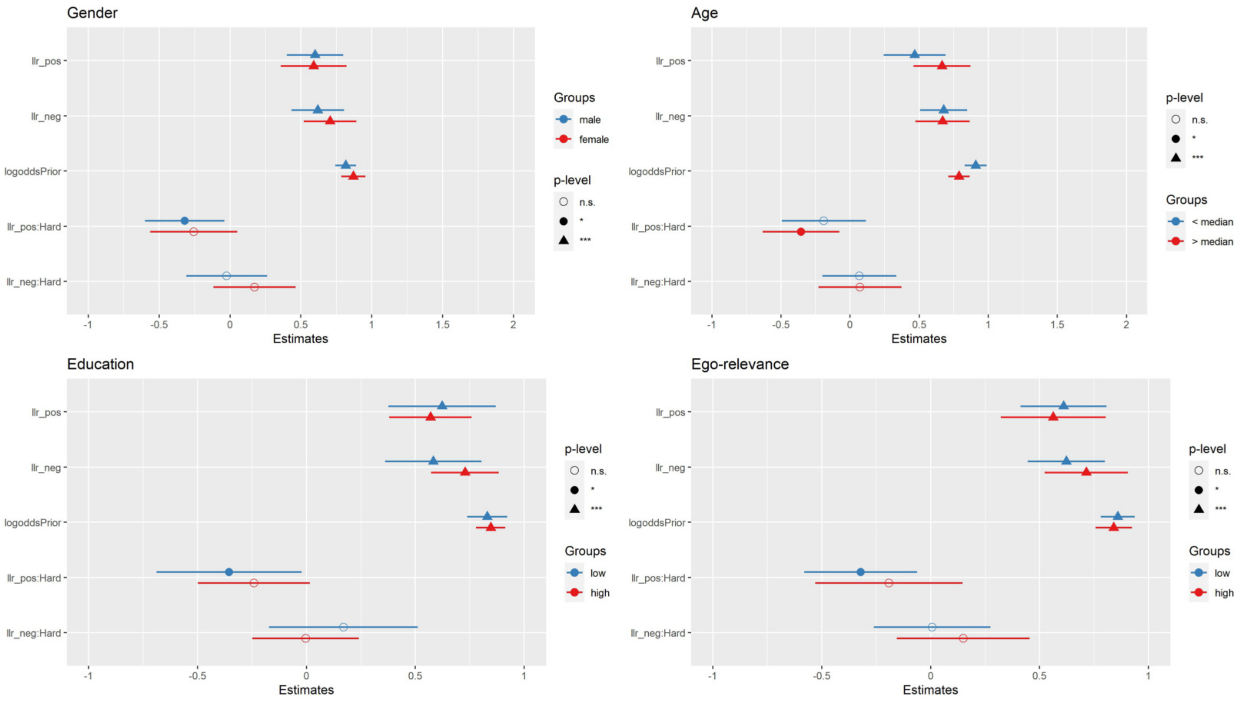

Figure A2: Heterogeneous belief updating? Subgroup analysis by various covariates

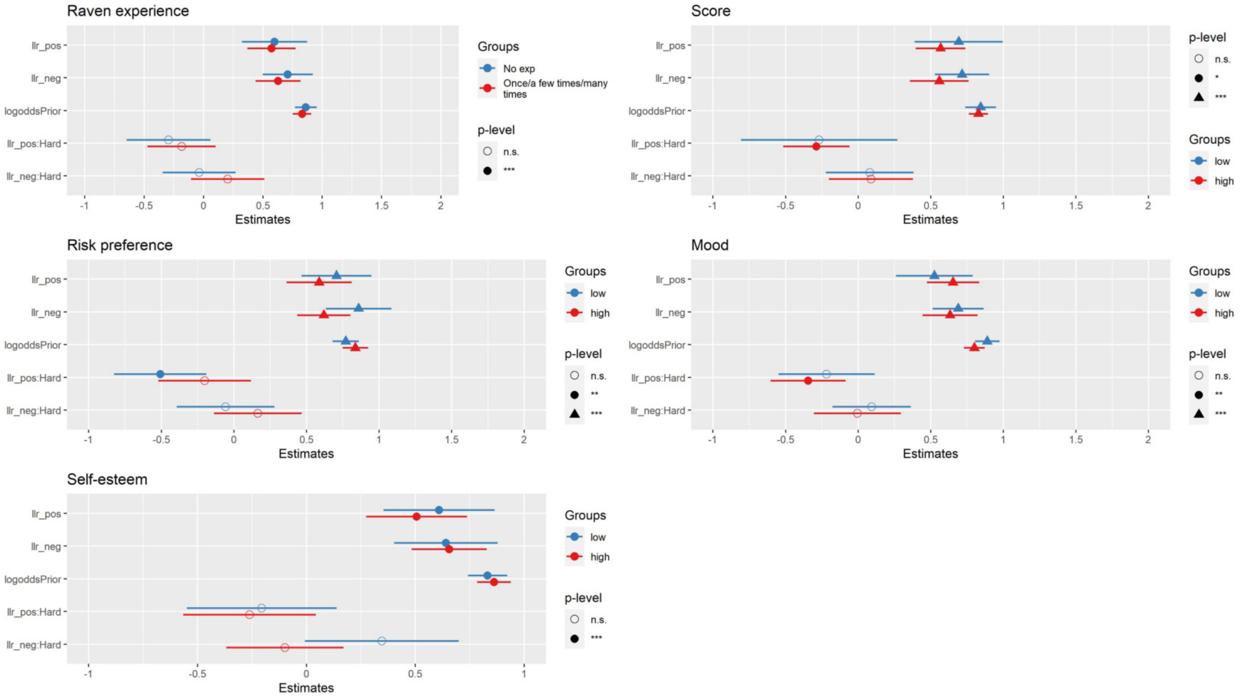

Figure A3: Heterogeneous belief updating? Subgroup analysis by various covariates

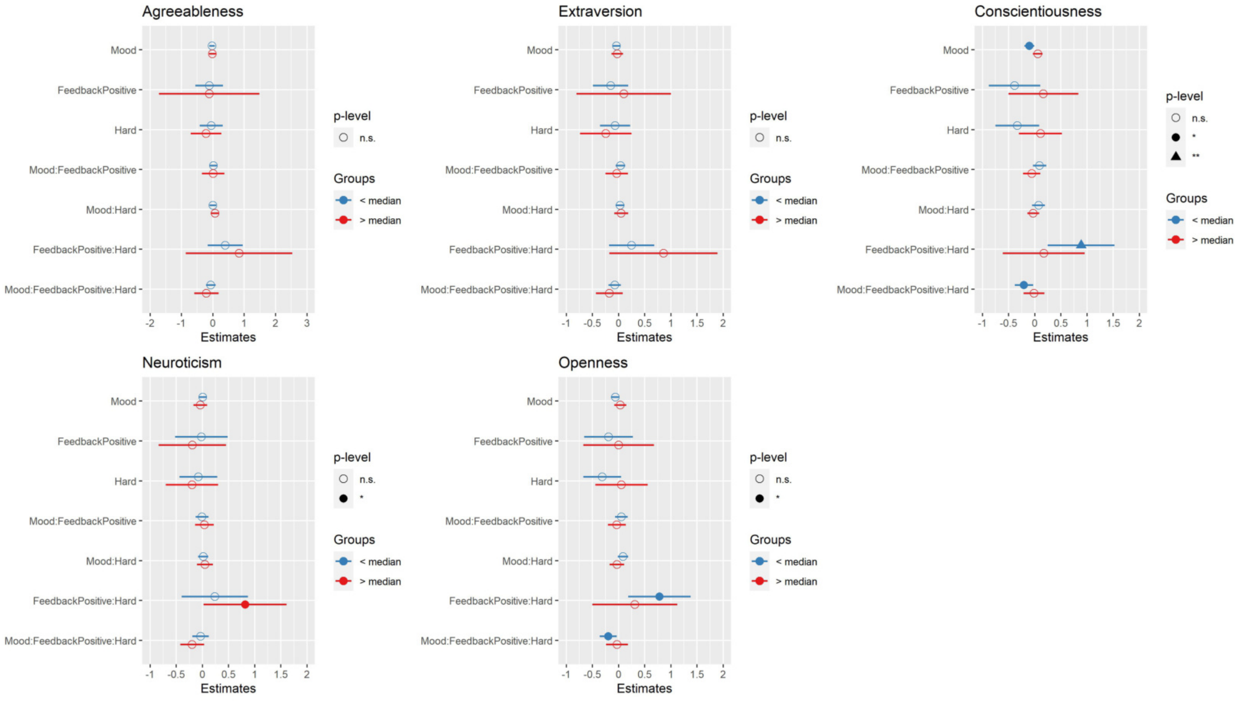

Figure A4: Heterogeneous belief updating? Subgroup analysis by Big 5

A.3.4 Subgroup analysis: Recall

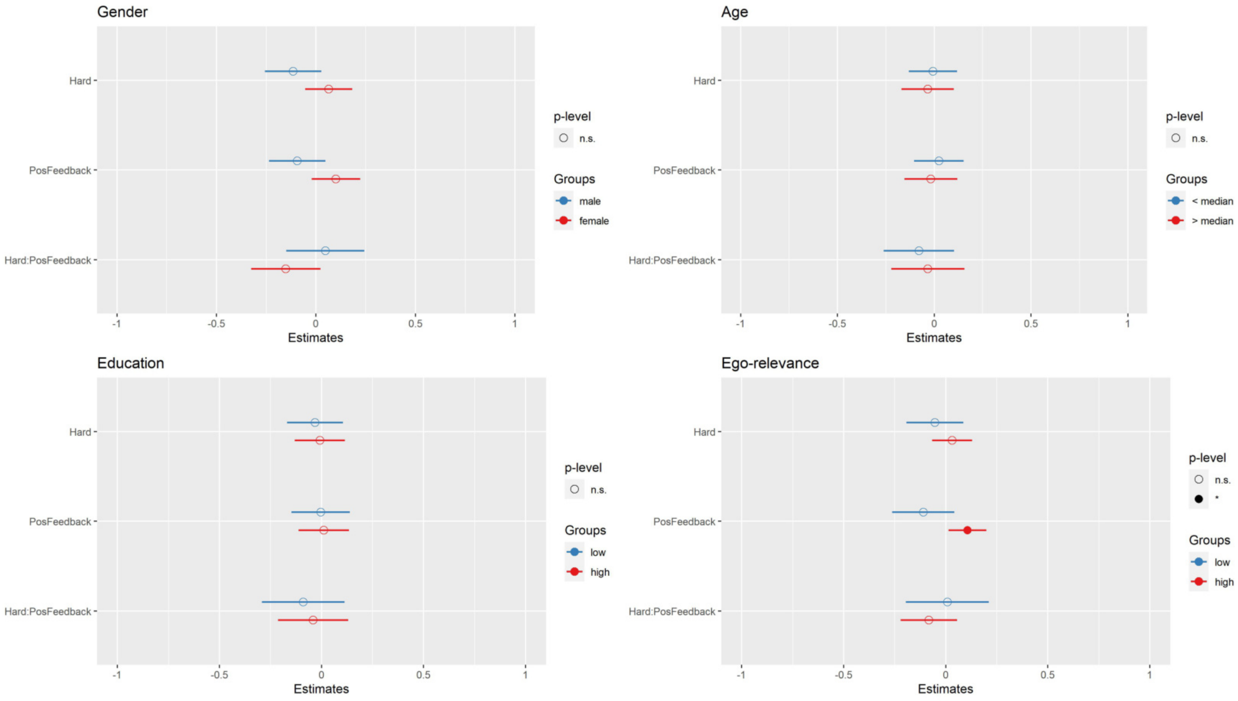

Figure A5: Heterogeneous recall? Subgroup analysis by various covariates

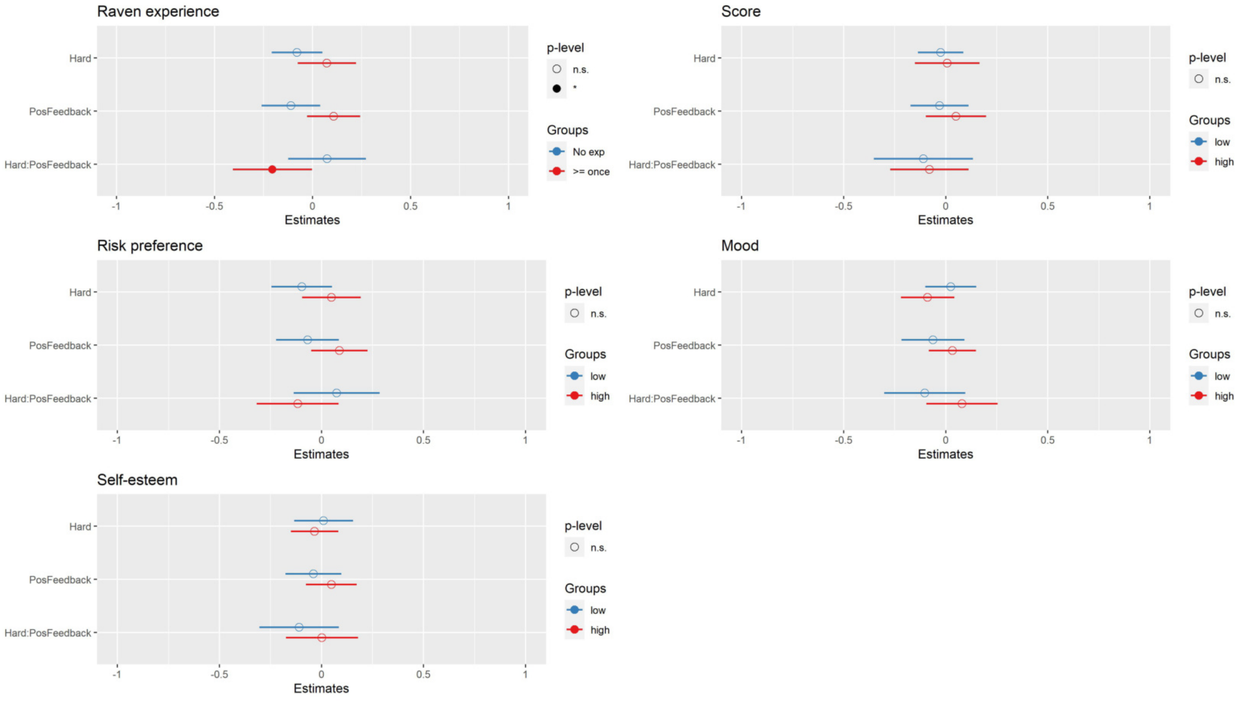

Figure A6: Heterogeneous recall? Subgroup analysis by various covariates

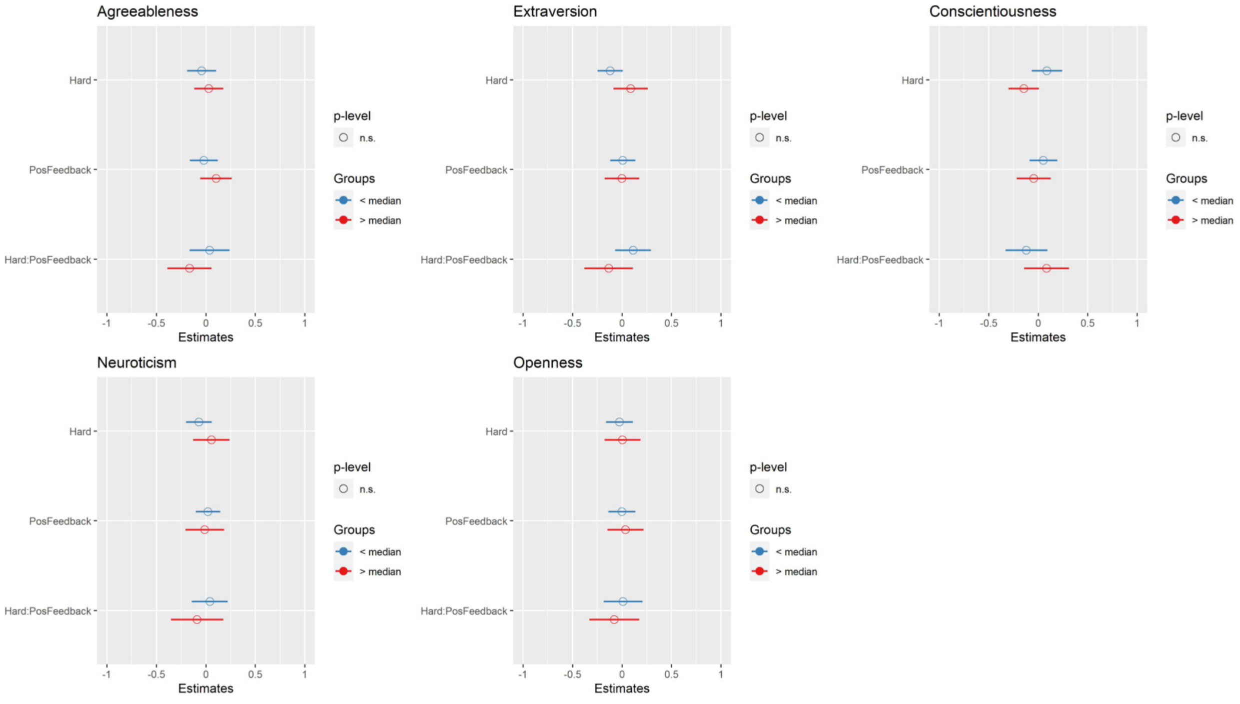

Figure A7: Heterogeneous recall? Subgroup analysis by Big 5