No evidence of biased updating in beliefs about absolute performance: A replication and generalization of Grossman and Owens (2012)

Quentin Cavalan, Vincent de Gardelle, Jean-Christophe Vergnaud

No evidence of biased updating in beliefs about absolute performance: A replication and generalization of Grossman and Owens (2012)

Quentin Cavalan [a, b, c, *]
Vincent de Gardelle [a, c, d]
Jean-Christophe Vergnaud [c, d]

a PSE, 48 Boulevard Jourdan, 75014 Paris, France
b UniversitéParis 1, 12 Place du Panthéon, 75231 Paris, France
c CES, 106-112, boulevard de l'Hôpital, 75013 Paris, France
d CNRS, 3 Rue Michel Ange, 75016 Paris, France

ABSTRACT

Many studies report that following feedback, individuals do not update their beliefs enough (a conservatism bias), and react more to good news than to bad news (an asymmetry bias), consistent with the idea of motivated beliefs. In the literature on conservatism and asymmetric updating, however, only one prior study focuses on judgments on absolute performance (Grossman & Owens, 2012), which finds that belief updating is well described by the Bayesian benchmark in that case. Here, we set out to test the replicability of these results and their robustness across several experimental manipulations, varying the uncertainty of participants' priors, the tasks to perform, the format of beliefs and the elicitation rules used to incentivize these beliefs. We also introduce new measures of ego-relevance of these beliefs, and of the credibility of the feedback received by participants. Overall, we confirm across various experimental conditions that individuals exhibit no conservatism and asymmetry bias when they update their beliefs about their absolute performance. As in Grossman & Owens (2012), most observations are well-described by a Bayesian benchmark in our data. These results suggest a limit to the manifestation of motivated beliefs, and call for more research on the conditions under which biases in belief updating occur.

Keywords biased updating, conservatism, asymmetry, feedback

JEL C91,D83

This work was supported by the Centre d'Economie de la Sorbonne (AAP-2020) and by the Agence Nationale de la Recherche [ANR-18-CE28-0015]

Colors are not necessary in print

1. Introduction

Forming correct beliefs about our environment and about ourselves seems essential in order to adapt our behavior and select the best course of action. However, prior research in experimental economics and psychology has repeatedly documented that individuals exhibit various biases in their beliefs. For instance, many studies show that individuals exhibit overconfidence : they overestimate their abilities to perform a task, relative to their true objective performance or to others performing the same task (Hoffrage, 2017; Menkhoff et al., 2013; Moore & Healy, 2008; Svenson, 1981; West & Stanovich, 1997; Grieco & Hogarth, 2009). When new pieces of information arrive, individuals also show biases in how they adjust their beliefs. In particular, individuals show conservatism and adjust their beliefs insufficiently, in comparison to an ideal Bayesian benchmark (Edwards, 1968; Fischhoff & Beyth-Marom, 1983; Huck & Weizsäcker, 2002; Lichtenstein & Slovic, 1971; Peterson & Miller, 1965). This conservatism is also found to be more pronounced in the case of bad news, a phenomenon referred to as asymmetric updating of beliefs (Eil & Rao, 2011; Sharot, 2011; Charness & Dave, 2017; Möbius et al., 2022). As an example, in Eil & Rao (2011), participants conformed to Bayes' Rule when learning they were more attractive than expected (good news) but exhibited conservatism when receiving negative feedback about their attractiveness (bad news). A related phenomenon is the observation that participants are more inclined to update their beliefs in the direction of new information, when this information confirms their prior views (Lord et al, 1979; Ditto & Lopez, 1992).

Importantly, although overconfidence and biases in belief updating have been documented in many studies as indicated above, it would not be correct to state that they are found in every individual in every situation. For instance, some studies also report no evidence for biased belief updating (Barron, 2021; Grossman & Owens, 2012) and overconfidence is known to depend on the base level of performance (Kruger & Dunning, 1999). Examining the conditions under which these biases arise thus remains an important empirical and theoretical question.

The survival of these biases, which might *a priori* seem unadaptive, have prompted researchers to investigate the possible benefits they might entail to individuals (Cazé & van der Meer, 2013). In this perspective, whether overconfidence and asymmetric belief updating are related phenomena has also been investigated (Benabou & Tirole, 2002; Benoît & Dubra, 2011; Buser et al., 2018). For instance, discarding bad news about oneself may help maintaining a positive self-image (Köszegi, 2006; Möbius et al., 2022), which in turn may keep individuals motivated to act and to engage in projects rather than stay idle (Benabou & Tirole, 2002; Compte & Postlewaite, 2004; Van den Steen, 2004). Ultimately, these biases may contribute to mental and physical health (Scheier et al., 1989; Strunk et al., 2006; Taylor et al., 2000).

Along these lines, it would be expected from a theoretical perspective that discarding bad news would be more pronounced for news relevant to the self-image of individuals. However, empirical investigations on this question have provided mixed results so far. On the one hand, several studies have found greater updating asymmetry (with greater reaction to good than to bad news) for ego-relevant information than for neutral information (Eil & Rao, 2011; Möbius et al., 2022). On the other hand, ego-relevance was not found to modulate belief updating in other studies (Buser et al., 2018; Coutts, 2019; Grossman & Owens, 2012). So far, the reasons for these discrepant results are still unclear. We note that experimental studies have mostly investigated contexts involving social comparisons between individuals, where the beliefs to be updated are about one's performance relative to another individual (was I better or worse than others in that task ?). To the best of our knowledge, a thorough experimental investigation of belief updating in the case of individuals who would learn about their own performance is still lacking. This stands in contrast with the situation considered by theoretical

models for asymmetric belief updating and overconfidence, which does not involve social comparisons (Benabou & Tirole, 2002; Compte & Postlewaite, 2004; Möbius et al., 2022).

One notable exception is the study by Grossman and Owens (2012) -thereafter G&O- who investigated beliefs about absolute performance (how many correct answers did I produce in this task?). In their study, G&O asked participants to answer 10 quiz questions, and to provide estimations about their performance before and after receiving noisy feedback. Specifically, participants reported the full distribution of probability over the possible scores, which allowed the authors to determine a Bayesian benchmark regarding belief revision. They found that a large majority of participants (around 70%) were well described by a the Bayesian benchmark, ego-relevant or not. Ego-relevant asymmetric updating was reported only in a subset of participants. However, as this study remains to date the only one investigating the updating of beliefs about absolute performance, a replication and generalization of these findings would help improve our understanding of biased belief updating.

In the present study, we set out to test the replicability of G&O results and their robustness across several experimental manipulations. In particular, we test our participants on two different tasks, a quiz task as in G&O, and a perceptual task where they simply have to evaluate whether an array of diamonds contains more empty or filled diamonds. This allows us to evaluate biased updating of beliefs about performance in two different domains (knowledge vs. perception) on which most past studies about updating were conducted, with tasks that can also differ in their base level of performance. In addition, since belief updating should depend on the uncertainty surrounding prior beliefs (the dispersion of the prior distribution), we manipulate this uncertainty by using two different levels of variability in the difficulty over items (i.e. the 10 quiz questions or the 10 perceptual trials). Specifically, in the low variability condition, the 10 items are all moderately difficult, and consequently participants' estimation of their performance should be more uncertain. In the high variability condition, by contrast, half of the items are very difficult but the other half involve very easy items. In this condition, the probability distribution over possible scores should have a smaller dispersion. Under the assumption that uncertain priors would be harder to update (because one has to update the likelihood of each probable score) and would lead to more conservatism and/or asymmetry, we expect that such biases would be more pronounced in the low variability condition. In addition to these task and variability manipulations, which took place in each participant, across participants we used two different elicitation rules for belief incentivization : a quadratic scoring rule as in G&O and a probability matching mechanism that incentivizes both performance and beliefs, and that is not sensitive to risk aversion. The rationale for this manipulation is that the quadratic scoring rule could in principle disincentivize performance in the task (see section 3.1.1 for more details) thus reducing the ego-relevance in this condition, which may ultimately lead to less asymmetry in belief updating. Finally, we also used two different types of estimations across participants: a full distribution of probabilities as in G&O vs. only the mean and a confidence interval, maybe more intuitive to participants.

To anticipate our results, our main contribution is to replicate G&O's findings across 8 experimental conditions (2 variability conditions, 2 types of tasks and 2 types of elicitation rules). First, the vast majority of participants exhibit neither conservatism nor asymmetry when they update beliefs about absolute performance. Second, we do not find any evidence for ego-relevant biased belief updating. Indeed, we reach similar conclusions when the task is deemed as more or as less ego-relevant by the participant (in our main study), and also similar conclusions when participants update beliefs about others (using a methodology more aligned with G&O, in a supplementary study). Another major contribution of this paper is to provide evidence that G&O's result is not driven by the belief format, as we do not find greater reaction to good news compared to bad news when participants provide mean estimates of their performance. Finally, using an original feature of our design, we show that participants are able to form consistent estimates of luck (i.e. whether they received feedback that was positively or negatively biased), although they exhibit some conservatism in this case.

2. Methods

2.1. Experimental Design

The experiment was run online. We recruited 305 participants on the Prolific platform. The experiment lasted around 60 minutes. Participants gained 9.5 pounds on average (around 13 dollars), which includes a show-up fee of 3 pounds (around 4 dollars).

After clicking on the experiment's link on Prolific, participants could read the instructions (see Supplementary Materials for the detailed instructions), and had to answer 3 comprehension questions. Participants who failed to answer correctly could not continue the experiment, and participants who answered correctly then proceeded to the main experiment.[1] The main experiment was made of 4 rounds. In each round, participants first performed a task, then they were asked to state their beliefs about their performance at the task before they received feedback about their performance (we refer to these beliefs as *prior* beliefs). After feedback, participants were asked to report their *posterior* beliefs about their performance. Finally, they were asked to indicate their beliefs about the type of feedback they received. Details about the procedure and measures are presented below.

2.1.1. Tasks

All participants performed 2 quiz tasks and 2 perceptual tasks. Across the 4 rounds, participants always performed the quiz tasks before the perceptual tasks. Each of the 2 quizzes consisted of ten multiple-choice questions selected from G&O's material. Those questions were selected from a book of Mensa quizzes which is designed to test IQ. In practice, those are standard logic questions in the same spirit as Mobius et al. (2022). Participants answered the questions one after the other and had to select one answer from a menu of 5 possible answers. The list of questions in each quiz is presented in the Supplementary materials. For the perceptual tasks, participants also performed 10 trials, where in each trial they had to indicate whether a 20x10 array of diamonds contained more empty or more filled diamonds. Each array was presented on the screen for 1 second.

To incentivize effort, for each task we selected one item (a question for the quiz task, a trial in the perceptual task) at random and participants earned 1000 points if their answer to this item was correct.

2.1.2. Variability manipulation

Within each type of task, the two sets of items differed in terms of the variability of the difficulty of the items in the set. In the *Low variability* condition, the items were of similar difficulty, whereas in the *High variability* condition, half of the items were very hard and half were very easy. The order of the *Variability* conditions was randomized between participants: 160 participants started with the *High* variability quiz first and 145 with the *Low variability* quiz first.

[1] Note that this questionnaire does not constitute a training where participants would learn to update their beliefs in a Bayesian way. In particular, it contains no feedback about the correct answer, and it is completely neutral with respect to the potential presence of asymmetry in belief updating.

For the quiz task, the *Low* and *High variability* quizzes were constructed by making participants in a pilot study answer a set of 30 questions and assess their probability to be correct after each question. Then, out of those 30 questions, we constructed 2 sets of 10 questions each with the constraint that the variability of subjective estimates between those two sets should be as far as possible and that the objective performance as well as the subjective probability estimates should be as close as possible (see Figure 1).

For the perceptual task, the stimulus difficulty was the difference between the number of filled and empty diamonds in the array. From the same pilot study, we estimated that differences of 2, 15, and 40 would lead to 50%, 75% and 100% of correct answers, respectively. Labeling these difficulty levels as low, intermediate and high, we constructed two types of sets: the *Low variability* set consisted of 10 trials at the intermediate difficulty level, while the *High variability* set consisted of 5 trials at the high difficulty level and 5 trials at the low difficulty level.

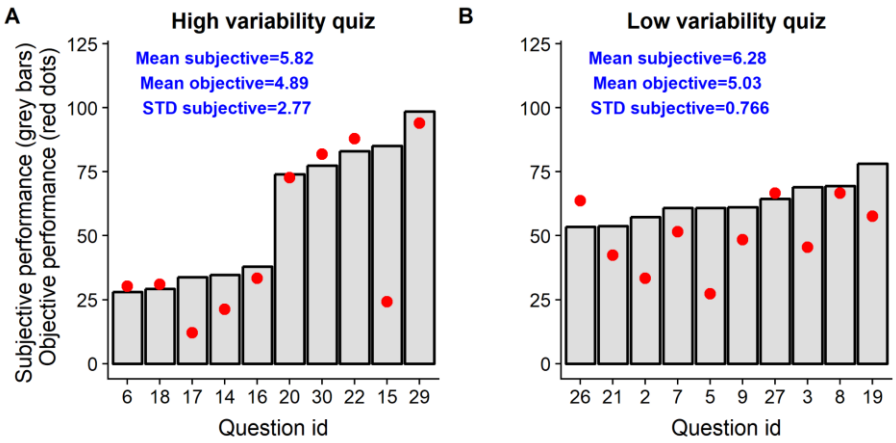

Figure 1. Average subjective performance estimates and objective accuracy for each question in the High variability quiz (panel A) and Low variability quiz (panel B), measured in a pilot session (N=80). Mean subjective estimate and objective accuracy across questions as well as standard deviation of subjective estimates are also reported in blue for each quiz.

2.1.3. Formats for performance belief elicitation

The format for the elicitation of prior and posterior beliefs was manipulated across participants.

In the *Distribution* condition, participants (N=152) were asked about the entire distribution of their beliefs, that is, for each possible score, the likelihood for this score (as in G&O). The sum of probabilities needed to be 100% before the participant was allowed to save the estimates.

In the *Mean* condition, participants (N=153) were asked to guess how many questions they answered correctly and to assess how sure they were of their estimate by giving an interval in which their actual score was included (the elicitation of the interval was not incentivized).

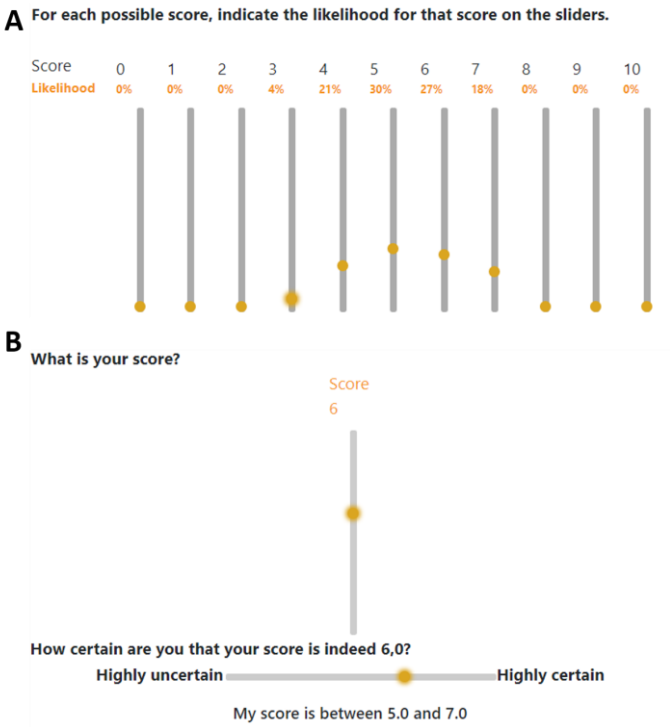

Figure 2. A. Screenshot of the elicitation of beliefs in condition Distribution. B. Same but in condition Mean

2.1.4. Elicitation rules for performance beliefs

Individuals' beliefs about absolute performance were incentivized via a Quadratic scoring rule as in G&O in the *QSR* condition (155 participants) or via a Becker-DeGroot-Marschak mechanism (Becker & DeGroot, 1974) in the *BDM* condition (150 participants).

In the QSR condition, participants could get between 0 and 1000 points for their beliefs. In the *Mean* condition, participants earned $1000 - 1000 \times \frac{(s-s^*)^2}{10}$, where s^* is the participant's actual score and s participant's estimate of his score. In the *Distribution* condition, participants earned $500 + 1000 \times p^* - w$ points, where p^* is the participant's estimated likelihood for their actual score and w is the sum of squares of the likelihoods for each of the eleven possible scores. Following Harrison et al. (2017), when reporting their beliefs in this condition, participants could see how many points they would get depending on their stated beliefs and their actual score.

In the *BDM* condition, participants' earnings are binary: participants could get either nothing or a reward of 1000 points. The mechanism depends on the belief format. In the *Mean* condition, one item in the set is randomly drawn. Then, a random number q is drawn from $[0,1]$. If q is smaller than $\frac{s}{10}$, participant earns 1000 points if the answer is correct and 0 otherwise. If q is higher, he earns 1000 points with probability q and 0 with probability $1-q$. In the *Distribution* condition, a number N between 0 and 10 is randomly drawn. Then, participant's belief that his score is higher or equal to N is recovered from the likelihood he gave for each score : $P(s^* \geq N) = \sum_{i=N}^{10} P(s^* = i)$. Then, a random number q is drawn from $[0,1]$. If q is smaller than $P(s^* \geq N)$, the participant earns 1000 points if his score is indeed higher or equal to N and 0 otherwise. If q is higher, he earns 1000 points with probability q and 0 with probability $1-q$. When reporting their beliefs in the *Distribution* condition, participants could see how the distribution they reported translated in terms of beliefs about $P(s^* \geq N)$.

A risk neutral participant will maximize expected utility in *QSR* when he reports truthfully his subjective expected number of correct answers (in the *Mean* condition) or his subjective probability distribution over scores (in the *Distribution* condition). In *BDM*, this is the case regardless of attitude towards risk (the mechanism is similar to Qu (2012) in the *Distribution* condition, see Supplementary Materials for the proof).

Finally, participants were told that the rule could appear complicated but that they would maximize their probability to win the reward (in *BDM*) or their average number of points (in *QSR)* by reporting their best guess.

2.1.5. Feedback

After reporting their first score estimate, participants receive noisy but unbiased feedback about their actual score. Following Möbius et al. (2022), subjects are told that a "hint" about their score is given to them by one of three different Martians: Inflated, Wise and Deflated Bob. Inflated and Deflated Bob each have 25% chance to be chosen and Wise Bob has 50% chance to be chosen. If selected, Deflated Bob gives a feedback lower than the actual score, by 1 or 2 units with probability 60% and 40% respectively. Inflated Bob gives a feedback higher than the actual score, by 1 or 2 units with probability 60% and 40% respectively. Wise Bob always gives the actual score. On top of Figure 2 which was presented to participants, we explained to them that feedback would be correct with 50% probability, inflated or deflated by 1 unit with 15% probability each and inflated or deflated by 2 units with 10% probability each, and that their score could not be more than 2 units away from the feedback.

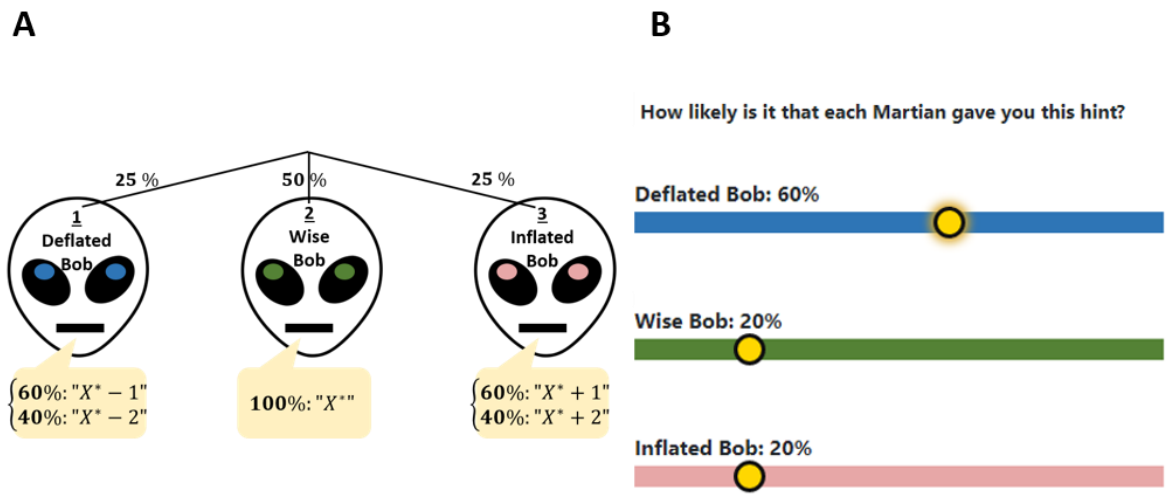

Figure 3. A. Structure of the feedback, as explained to participants in the instructions. X refers to the actual score of the participant. B. Screenshot of the elicitation of belief about luck.*

2.1.6. Beliefs about Luck

After giving their posterior belief about their score, participants had to indicate for each possible Martian their belief that this Martian actually gave the feedback. We call those beliefs "beliefs about luck" as it captures whether participants believe the feedback they receive was unluckily below their actual score (Deflated Bob gave it), correct (Wise Bob gave it) or luckily above their actual score (Inflated Bob gave it). Again, participants' earnings for their estimates depended on the elicitation rule condition: *QSR* or *BDM*.

In the QSR condition, participants earned $500 + 1000 \times p^* - w$ points, where p^* is the participant's estimated likelihood for the Martian that was actually chosen and w is the sum of squares of the likelihoods for each of the 3 possible Martians. In the BDM condition, beliefs about luck are paid similarly to beliefs about performance. One Martian is randomly drawn. Then, a random number q is drawn from $[0,1]$. If q is smaller than the participant's belief that this Martian was chosen, he earns 1000 points if this Martian was indeed chosen and 0 otherwise. If q is higher, he earns 1000 points with probability q and 0 with probability $1 - q$. Again, *QSR* ensures truthful reporting only if participants are risk neutral while *BDM* do so regardless of risk attitudes.

2.1.7. Relevance of the task

Before receiving their payment, participants were asked one question which we used as a proxy measure for the ego-relevance of the tasks. In practice, we asked them: "Which task is the most representative of skills you value in everyday life?". They could answer "both tasks", "the perceptual task", "the quiz task" or "neither task". Their answer was not incentivized.

2.1.8. Payment

In each round, participants could earn points at four different stages (through the quiz/perceptual tasks, through prior beliefs, through posterior beliefs and through luck beliefs). At the end of the experiment, two rounds were randomly drawn, one for the quiz part and one for the perceptual part.[2] Then, one stage was randomly drawn in each part and the points earned at this stage were converted into pounds with 1000 points = 6 pounds (8$).

2.2 Measures

2.2.1. Mean beliefs and mean biases

In the *Distribution* condition, we recover from participants $(Prior_i)_{i\in[\![0,10]\!]} = P(score = i)_{i\in[\![0,10]\!]}$ and $(Post_i)_{i\in[\![0,10]\!]} = P(score = i|feedback)_{[\![0,10]\!]}$ that is participants' prior and posterior belief about their score.

We define the mean prior and posterior beliefs as:

$$Mean\ Prior = \sum_{i=0}^{10} i \times Prior_i$$

$$Mean\ Post = \sum_{i=0}^{10} i \times Post_i$$

In the *Mean* condition, we directly ask mean prior and posterior beliefs to participants.

By comparing the predicted score and actual score, we define prior bias and posterior bias as:

$$Prior\ bias = Mean\ Prior - Score$$

[2] This means there could potentially be hedging across quiz and perceptual tasks.

$$Post\ bias = \ Mean\ Post - \ Score$$

2.2.2. Uncertainty of beliefs

We use Shannon's entropy (Shannon, 1948) to measure participant's level of uncertainty regarding their performance beliefs.

$$Prior\ Entropy = -\sum_{i=0}^{10} \ Prior_i \times log(Prior_i)$$

$$Post\ Entropy = -\sum_{i=0}^{10} \ Post_i \times log(Post_i)$$

The advantage of entropy is that it provides a measure of the dispersion of a probability distribution independent of the mean.

2.2.3. Bayesian posterior distribution about score

In the *Distribution* condition, our design allows us to construct a Bayesian benchmark for posterior performance beliefs given prior beliefs. More precisely, an ideal Bayesian agent with prior beliefs about his score $(Prior_i)_{i\in[\![0,10]\!]}$ who receives feedback f should form posterior beliefs $(Post_i^*)_{i\in[\![0,10]\!]}$defined as

$$Post_i^* = P^*(score = i|f) = P(f|i) \times \frac{Prior_i}{P(f)}$$

$$where\ P(f|i) = \begin{cases} .1\ if\ \ |i-f|=2 \\ .15\ if\ \ |i-f|=1 \\ .5\ if\ \ |i-f|=0 \\ 0\ otherwise \end{cases} \quad and\ \ P(f) = \sum_{j=0}^{10} P(f|j) \times Prior_j$$

Note that when the prior distribution only assigns positive probability to scores outside $[\![f-2, f+2]\!]$ then, $P(f) = 0$ and thus Bayes' Rule provides no benchmark for the posterior distribution. Following G&O, we construct a Bayesian posterior proxy in this case. In a nutshell, we assign probability 1 to the "closest" score relative to participant's mean prior belief compatible with the feedback. More precisely, if $Mean\ prior > f$, we set $Post_i^* = 1$ if $i = f + 2$ and 0 otherwise. If $Mean\ Prior < f$, we set $Post_i^* = 1$ if $i = f - 2$ and 0 otherwise.

Note that such a Bayesian benchmark cannot be constructed in the *Mean* condition.

2.2.4. Perceived and Bayesian luck

In the *Distribution* condition, we also construct a measure which aggregates the distribution of beliefs about Martians into a single number, called perceived luck. Perceived luck corresponds to the participant's expectation of the noise in the feedback he received. Because feedback is unbiased, before receiving it, expected noise is 0. Recall that noise added to the feedback can be either -2 or -1 (Deflated Bob), 0 (Wise Bob), +1 or +2 (Inflated Bob). Thus, denoting by $b_{\epsilon=i}$the belief that a particular noise sample ϵ was realized given the feedback, i.e. $b_{\epsilon=i} = P(\epsilon = i|feedback = f)$, the beliefs about each Martian correspond to $b_{\epsilon>0}$, $b_{\epsilon<0}$, and $b_{\epsilon=0}$.

In order to compute expected noise given the feedback, we are lacking two estimates that would indicate the relative probability of a noise of 2 vs a noise of 1 given that noise is positive, that is $b_{\epsilon=2|\epsilon>0}$, and the relative probability of a noise of -2 vs a noise of -1 given that noise is negative, that is $b_{\epsilon=-2|\epsilon<0}$. We use their Bayesian counterparts as a proxy for these probabilities. In the *Distribution* condition, we can indeed construct the Bayesian benchmark for the probability that a particular noise sample ϵ was realized given feedback f, using the Bayesian posterior belief about performance defined before:

$$b^*_{\epsilon=i} = Post^*_{f-i}\ if\ f-i \geq 0\ and\ f-i\ \leq 10, 0\ otherwise$$

We can then compute $b^*_{\epsilon=2|\epsilon>0} = \frac{b^*_{\epsilon=2}}{b^*_{\epsilon=1}+b^*_{\epsilon=2}}$ and $b^*_{\epsilon=-2|\epsilon<0} = \frac{b^*_{\epsilon=-2}}{b^*_{\epsilon=-1}+b^*_{\epsilon=-2}}$ to define perceived luck as :

$$Perceived\ Luck = -b_{\epsilon<0} \times \ (b^*_{\epsilon=-2|\epsilon<0} + 1) + b_{\epsilon=0} + b_{\epsilon>0} \times (b^*_{\epsilon=2|\epsilon>0} + 1)$$

While Bayesian luck is defined as:

$$Bayesian\ luck = \sum_{i\in[-2,2]}\ \ i\ \times b^*_{\epsilon=i}$$

Note that using the Bayesian benchmark to proxy some probabilities determining perceived luck means that the level of biases we find with this measure will be a lower bound. To address this issue, we show in Supplementary Materials that our results are robust to the use of a proxy for $b_{\epsilon=-2|\epsilon<0}$ and $b_{\epsilon=+2|\epsilon>0}$ which rely on participant's posterior distribution instead of the Bayesian posterior.

2.3. Main differences with G&O

Here, we clarify some differences between our design and the design of G&O. Most of them are actually related to the main contribution of this paper, that is the manipulation of 4 experimental factors not present in G&O (the type of task, the uncertainty of beliefs, the type of elicitation rule, and the type of belief format). Moreover, we introduce a new measure of perceived luck, and our operationalization of ego-relevance is different compared to G&O. Participants' beliefs about luck are indirectly inferred from the difference between feedback and mean posterior beliefs in G&O. In our experiment, we directly ask participants such beliefs which allows us to evaluate separately belief biases on luck and on performance. Another important discrepancy between the two designs is that all participants in our study are asked about their own performance while in G&O, some participants are actually asked about the performance of another participant. The authors use this manipulation to compare updating when it is ego-relevant vs when it is not. We address this question using a different methodology. In practice, we ask participants which task they think relates to valuable skills which enables us to compare updating when the task is considered relevant vs when it is not. However, for a proper comparison with G&O, we

also measured the updating of beliefs about the performance of another participant, in a supplementary study (presented in supplementary material and summarized in our discussion).

Our design also departs from G&O's in two less critical aspects. First, we do not ask performance beliefs to participants before doing the task. This means we do not study the way participants update their beliefs after actually experiencing the task. This is because our goal was to evaluate biases in belief updating and that such investigation was impossible in this case, due to the lack of well-defined Bayesian benchmark. Second, although in both studies the noise of the feedback is drawn between $[\![f-2, f+2]\!]$ and imposes a truncation on the posterior between $f-2$ and $f+2$, it is less informative in G&O. Indeed, it is uniformly distributed in their study while in our's, it is more likely to be 0 (50%) than 1 or -1 (15% each) than 2 or -2 (10% each). We made this change because a potential explanation for the absence of updating biases in G&O could have been that Bayes' Rule was actually too "simple" to compute in their setting, leaving no room for biases to emerge. Indeed, in G&O, Bayes' rule simply prescribes that posterior belief should be truncated between $[\![f-2, f+2]\!]$ and within this range, in the same proportions as in prior beliefs. In our setting, Bayes' rule prescribes that posterior scores within $[\![f-2, f+2]\!]$ should be reweighted according to their degree of proximity with the feedback which is arguably more complex.

3. Results

First, we start by replicating G&O's result that belief updating is overall unbiased (that is, neither conservatism nor asymmetric) across our 8 experimental conditions, whether beliefs are ego-relevant or not. Because this analysis requires a Bayesian benchmark to compare with participant's updating, we perform this analysis only in the *Distribution* condition. Second, comparing updating between *Distribution* and *Mean*, we provide evidence that the absence of asymmetry in our data and in G&O is not driven by the format of beliefs: beliefs do not appear to be updated more asymmetricaly (in the sense of greater reaction to good news than bad news) when participants report their expected number of correct answers (*Mean* condition) compared to a full distribution of probabilities (*Distribution* condition). Finally, we analyze beliefs about luck and show that although they exhibit some conservatism, they do not reveal asymmetry either.

3.1. Beliefs and performance in the *Distribution* condition

Overall, participants' average score is equal to 6.42 out of 10. Their average mean prior is 5.98 and their average mean posterior is 6.30. Thus, participants underestimate their scores both at the prior (PriorBias=-.43, t(151)=-5.46, p<.001) and posterior (PostBias=-.11, t(151)=-2.28, p=.024) stage. Consistent with this posterior underestimation, participants believe that the feedback they receive is lucky (PerceivedLuck=0.07, t(151)=2.51, p=.013). In addition, the entropy of beliefs decreases from the prior stage to the posterior stage (PriorEntropy=1.23, PostEntropy=1.0, t(151)=10.46, p<.001). We describe next how performance and beliefs depend on our experimental factors.

3.1.1. Elicitation rules do not affect participants' beliefs and performance

First, we perform one-way ANOVAs to assess the effect of the type of elicitation rule being used (BDM or QSR) on participants' beliefs and performance.

The first concern that justifies the comparison between BDM and QSR rules is that with QSR (but not with BDM) participants could strategically decrease their performance in the quiz/perceptual task in order to increase their earnings in the belief elicitation task.[3] However, we find no evidence of such behavior in our data: average scores are not significantly different between conditions (Score_BDM=6.44, Score_QSR=6.29, F(1,150)=0.075, p=.785).

The second concern is that QSR could alter reported beliefs because it is not incentive compatible under risk-aversion, while BDM is. Again, we find no evidence of such change in reported beliefs as we find no effect of rules on mean prior belief (F(1,150)=0.309, p=.579), mean posterior belief (F(1,150)=0.016, p=.901), prior belief entropy (F(1,150)=.341, p=.56), posterior belief entropy (F(1,150)=.068, p=.795) and perceived luck (F(1,150)=0.032, p=.858).

3.1.2. Effects of variability and task on beliefs and performance

Table 1 presents descriptive statistics for each outcome variable (scores, mean prior and posterior beliefs, belief biases, belief entropy and perceived luck), as a function of task and variability conditions in our experiment. Given that BDM and QSR produced similar results in terms of beliefs and performance, for simplicity in this section we conduct ANOVAs to test the effects of task and variability on each outcome variable, without including the type of scoring rule as a factor. These analyses show no interaction between task and variability for any of the outcome variables (all p>.05).

Within Factors		Outcome variables							
Task	**Variability**	Score *(1)*	Mean prior belief (2)	Prior bias (3)	Prior belief entropy (4)	Mean posterior belief (5)	Posterior bias (6)	Posterior belief entropy (7)	Perceived Luck (8)
Quiz	Low	5.38 (0.17)	5.05 (0.17)	-0.33 (0.15)	1.23 (0.04)	5.33 (0.17)	-0.05 (0.09)	1.06 (0.04)	0.07 (0.06)
Quiz	High	5.41 (0.16)	5.45 (0.15)	0.04 (0.14)	1.19 (0.04)	5.44 (0.15)	0.03 (0.09)	1.00 (0.04)	0.01 (0.06)
Perceptual	Low	7.38 (0.13)	6.44 (0.11)	-0.94 (0.16)	1.28 (0.04)	7.07 (0.12)	-0.31 (0.09)	1.05 (0.04)	0.19 (0.06)
Perceptual	High	7.49 (0.10)	7.00 (0.11)	-0.49 (0.14)	1.21 (0.04)	7.36 (0.11)	-0.13 (0.08)	1.04 (0.04)	0.01 (0.05)
Main effect of task		***	***	***	ns	***	*	ns	ns

[3] This is the case even for a risk neutral participant. For instance, a participant who knows the correct answer to most questions might purposely select incorrect answers to the remaining questions if what he gains from this strategy (an increase in QSR's expected earnings due to more precise beliefs) is greater than what he loses (a decrease in task's expected earnings due to lower score). This tradeoff depends on how task performance is rewarded compared to belief accuracy, as well as on the subject's uncertainty surrounding his performance.

Main effect of variability	ns	***	**	**	ns	ns	ns	*

*Table 1. Mean of scores, mean prior belief, prior bias, prior belief entropy, mean posterior belief, posterior bias, posterior belief entropy and perceived luck across participants by within-participant factor (type of task and variability) in Distribution condition. Standard deviations are reported in parenthesis. The last two lines report, for each outcome variable, the significance of the main effect of each factor in an ANOVAs with variability and type of task and their interaction as within participant factor (p<.05 *, p<.01 **, p<.001 ***)*

Regarding the main effect of the task, lower scores are found in the quiz task than in the Perceptual task, as expected from our difficulty calibration procedure. This difference in scores is accompanied by differences in beliefs, but also in belief estimation biases: with greater underestimation of scores in the Perceptual task, both for prior and posterior estimates. Belief entropy and perceived luck do not differ across tasks.

We now report the effects of the variability manipulation. Firstly, task scores are virtually identical between variability conditions. However, prior beliefs are higher in the *High variability* condition, resulting in larger underestimation in *Low variability* than in *High variability*. These effects are not present for posterior beliefs. Importantly, when examining entropy, we find as expected that prior beliefs are more uncertain (the distribution over scores is more dispersed) in the *Low variability* than in the *High variability* condition. This difference is not present anymore for posterior beliefs.

3.1.3. Performance belief updating is unbiased for most participants

To estimate biases in belief updating, we compare participants' actual updates (the difference between the mean posterior and the mean prior belief) to the update that would be computed by an ideal Bayesian agent using the same prior as participants and receiving the same feedback as they receive. To do so, we consider the following regression model:

$$Update_{Actual} = \alpha + \beta \times update_{Ideal} + c \times GN + \gamma \times update_{Ideal} \times GN \quad (1)$$

In eq. (1), GN is equal to 1 when the feedback is higher than the mean prior (a good news) and to 0 when it is lower (a bad news). β is the extent to which participants follow the ideal in the case of bad news, and γ is the good news vs. bad news effect on this ability to update beliefs optimally. Ideally, β should be equal to 1 (values below 1 indicate conservatism) and γ equal to 0 (values above 0 indicate asymmetric updating with stronger inference from good than bad news). In addition, α and c are constants of no interest.

Following the methodology of G&O, we estimate conservatism and asymmetry in belief updating by considering 3 different categories of observations: those for whom beliefs cannot be updated in a Bayesian manner because scores included in the prior distribution are not within the range of scores compatible with the feedback (*Not Updateable*), those for whom parts of the posterior distribution over scores are not within the truncated range imposed by the feedback (*Not Truncated*), and those who do not belong to the previous two categories (*Truncated*). Overall, we find that the vast majority of observations (83%) belong to the *Truncated* category, whereas 6% are *Not Updateable* and 11% are *Not Truncated.* These proportions do not change across our experimental conditions (X-squared(14)=18.18, p=.199), and they are roughly comparable to those in G&O, who reported more *Not Updateable* (13%) and fewer *Truncated* observations (72%).

Figure 4 illustrates the updating of beliefs for these 3 categories of observations, in our different experimental conditions. Critically, we find that in all conditions, most observations fall along the diagonal, indicating close alignment to Bayesian updating. Only a few observations in the *Not Truncated*

category seem to depart from Bayesian updating. We thus find no sign of conservatism or asymmetry in most observations, in all conditions. This replicates the main finding of G&O.

Table 2 reports the coefficient estimates of the regression on updates (eq (1)). As in G&O, we find that when all observations are included in this regression, there is some evidence for conservatism and asymmetry (here, in two of the quiz conditions out of 8 conditions). However, examining regressions based on observations excluding the *Not Truncated* or the *Not Updateable* suggests that biases are mainly due to the *Not Truncated* category (See Table S1 in Supplementary Materials). When focusing on the *Truncated* observations, the evidence is in favor of Bayesian updating without conservatism or asymmetry. The only condition where participants depart significantly from Bayesian updating (column 6 in the table), in fact shows oversensitivity to feedback ($\beta = 1.51 > 1$), not conservatism.

To further evaluate the need to include conservatism and asymmetry parameters in the regression, we compare the BIC values between two regression models, one with both conservatism and asymmetry as free parameters, and one where these parameters are forced to their ideal values (β=1 and γ=0). For the *Truncated* observations, pooling across all conditions, the BIC value is lower for the most restricted model (BIC = 2986.58) than for the model with the two free parameters (BIC = 2994.56), indicating that the restricted model is better. The difference in BIC values of about 8 indicates strong support for the restricted model over the model with the two bias parameters (Burnham & Anderson, 2002).

Conditions	Variability	Pooled	High	Low	High	Low	High	Low	High	Low
	Task	Pooled	Quiz	Quiz	Quiz	Quiz	Percept	Percept	Percept	Percept
	Rule	Pooled (1)	BDM (2)	BDM (3)	QSR (4)	QSR (5)	BDM (6)	BDM (7)	QSR (8)	QSR (9)
All obs ervat.	β	0.70* (0.12)	0.33*** (0.21)	1.24 (0.15)	0.74 (0.15)	0.40*** (0.15)	0.97 (0.35)	1.29 (0.27)	1.08 (0.24)	0.92 (0.21)
	γ	0.08 (0.14)	0.21 (0.24)	-0.13 (0.20)	0.36 (0.26)	0.39 (0.22)	-0.22 (0.38)	-0.45 (0.30)	-0.07 (0.26)	-0.40 (0.24)
	c	8.53*** (1.44)	11.4*** (3.18)	2.10 (2.28)	5.93* (2.88)	8.29** (2.14)	7.73* (3.49)	4.02 (3.10)	7.53** (2.20)	11.50** (3.76)
	a	-4.06*** (0.78)	-6.77** (2.10)	-0.08 (1.64)	-3.87* (1.87)	-3.72 (1.90)	-3.05 (2.75)	-1.08 (2.54)	-2.91 (1.70)	-3.27 (3.12)
	R^2	.68	.52	.81	.67	.64	.64	.69	.85	.65
	N	589	75	76	73	72	74	74	73	72
Only tru nc	β	1.11 (0.08)	1.32 (0.21)	1.42 (0.26)	0.98 (0.14)	0.89 (0.14)	1.51** (0.19)	1.27 (0.14)	1.24 (0.24)	0.87 (0.37)
	γ	-0.07 (0.08)	-0.03 (0.27)	-0.35 (0.30)	0.34 (0.21)	-0.01 (0.19)	-0.38 (0.23)	0.03 (0.18)	-0.28 (0.27)	0.07 (0.38)
	c	5.20*** (0.77)	3.00 (2.03)	2.91 (1.99)	2.58 (1.99)	8.13*** (2.14)	3.33 (1.93)	1.30 (1.65)	6.58** (2.16)	7.83** (2.48)
	a	-2.11*** (0.48)	-1.66 (1.41)	-0.07 (1.55)	-1.74 (1.32)	-3.57 (1.60)	0.16 (1.43)	-0.97 (1.27)	-2.51 (1.68)	-3.49 (2.20)

	R^2	.85	.83	.81	.82	.83	.88	.90	.84	.86
	N	484	63	58	62	55	59	62	67	58

Table 2. Regressions examining the conservatism and symmetry of the response to feedback by experimental condition, for all observations and for the subsample of Truncated observations only. Participants for whom feedback = mean prior belief are not included in this analysis, as it is considered neither good nor bad news. Asterisks applied to β denote significant difference from one. Asterisks on all other coefficients denote significant difference from zero. Standard errors in parentheses.

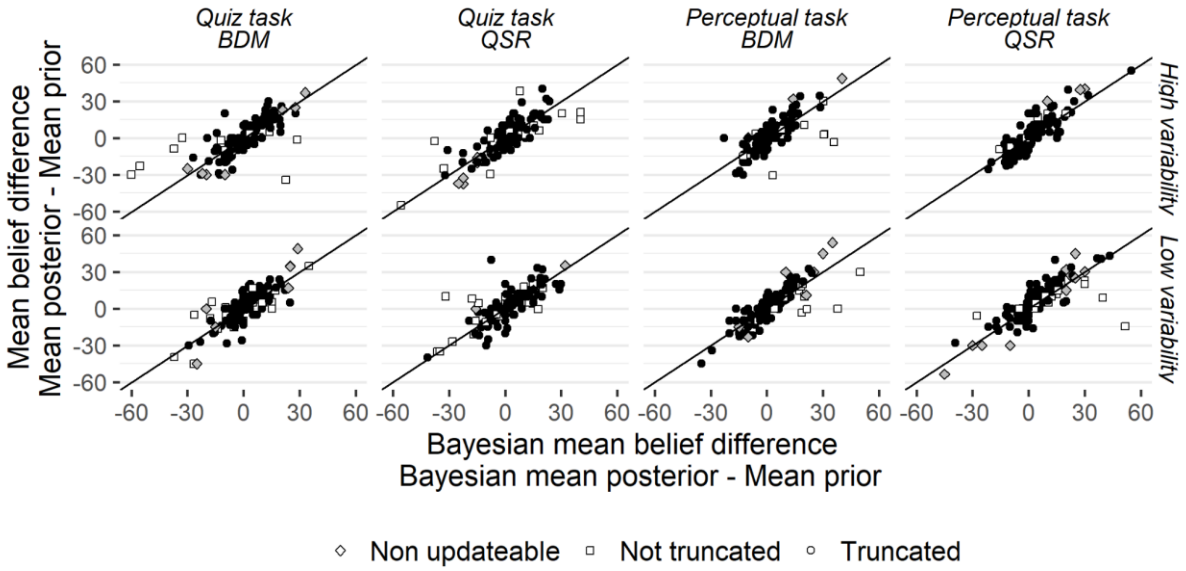

Figure 4. Actual updating (mean posterior beliefs - mean prior beliefs) versus Bayesian updating (Bayesian mean posterior - mean prior) for each experimental condition within condition Distribution.

3.1.4. No evidence for motivated belief updating

In G&O, the authors compare updating between two conditions where subjects are asked about beliefs regarding their own performance or the performance of another subject. If subjects derive utility from believing they have high performances, they could exhibit more asymmetry when updating beliefs about themselves than about others. However, in most observations (the *Truncated* category), G&O find that participants are close to Bayesian updating irrespective of whether beliefs are ego-relevant or not, and conclude for the absence of motivated belief updating in their setting.

In the present study, participants always update beliefs about their own performance, but in addition they indicate at the end of the experiment which task (the quiz or the perceptual task) was "most representative of the skills they value in daily life". In this section, we take this variable as a proxy for ego-relevance, and we focus more specifically on the participants who answered either the quiz task (65 participants) or the perceptual task (38 participants), excluding 40 participants who answered "both" and 9 who answered "neither".[4] Then, we evaluate how conservatism and asymmetry might be more pronounced for tasks deemed as more ego-relevant.

[4] Since participants answer this question at the end of the experiment, our measure might suffer from endogeneity problems (Drobner, 2022). To strengthen our claim that there is no evidence for motivated beliefs updating in our

Figure 5 and Table 3 indicate no evidence of conservatism and no evidence for under-reaction to bad news, in the ego-relevant or in the non-ego-relevant conditions. Including all observations or only observations in the *Truncated* category in our analysis produces the same results. In fact, contrary to what the presence of motivated beliefs would predict, participants show a larger sensitivity to bad news (β is numerically larger) in ego-relevant conditions. In addition, reactions to bad news are not more reduced compared to good news in the ego-relevant condition: on the contrary, γ is numerically more negative in the ego-relevant case. Overall, thus, the present data is not in line with what would be expected in the presence of motivated belief updating.

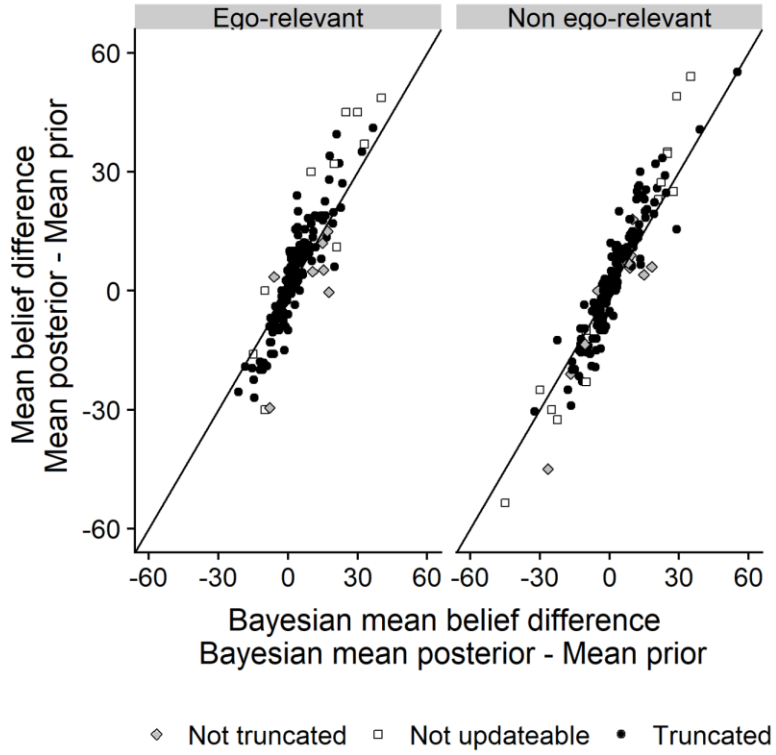

Figure 5. Actual updating (mean posterior beliefs - mean prior beliefs) versus Bayesian updating (Bayesian mean posterior - mean prior) depending on the ego-relevance of the task, within condition Distribution.

	All observations		Truncated only	
Condition	Ego (1)	Non ego (2)	Ego (3)	Non ego (4)
β	1.38** (0.14)	1.15* (0.08)	1.39** (0.14)	1.07 (0.12)
c	3.48** (1.04)	3.39*** (0.96)	3.95*** (1.04)	4.49*** (0.98)
γ	-0.31* (0.16)	-0.06 (0.12)	-0.35* (0.17)	-0.04 (0.14)

setting, in a supplementary study we therefore collected data replicating the condition “Other” in G&O. We go back to this issue in the discussion.

a	-1.16 (0.69)	-1.01 (0.62)	-1.21 (0.73)	-1.42 (0.76)
R^2	.83	.89	.86	.87
N	196	201	174	174

Table 3. Regressions examining the conservatism and symmetry of the response to feedback depending on the ego-relevance of the task, for all observations and for the subsample of Truncated observations only.

3.2. Updating mean vs. full beliefs about performance

One may think that participants do not exhibit conservatism and asymmetry in their belief updating because updating an entire distribution of belief over scores is unnatural and might promote deeper cognitive involvement. To address this issue, half of our participants had to indicate a full distribution of probabilities as in G&O, while the other half indicated their mean score and a confidence interval. In this section, we first report descriptive statistics about this *Mean* condition, showing that participants in this condition are slightly better at the task, and slightly less biased in their prior beliefs, compared to the *Distribution* condition. We then compare belief updating between these two conditions.

3.2.1. Beliefs about mean performance

Table 4 summarizes data on participants' scores and beliefs in the *Mean* and *Distribution* conditions, separately for the two tasks. The reader can refer to table S2 in Supplementary Materials for descriptions of performances and beliefs separated further by variability levels within the *Mean* condition. Scores are slightly higher in the *Mean* condition (F(1,303)=3.887, p=.050). This effect of format on scores interacts with tasks (F(1,303)=5.052, p=.025), as the format effect is present in the quiz task (Score_Mean=5.87, Score_Distrib=5.39, t(303)=2.410, p=.017) but not in the perceptual task (Score_Mean=7.44, Score_Distrib=7.42, t(303)=-0.205, p=.837). This may be due to the fact that belief elicitation is more demanding in the *Distribution* condition, which could disrupt participants' ability to perform the quiz, whereas the perceptual task, presumably more automatic, is not affected.

Interestingly, biases in the prior beliefs exhibit a main effect of belief format (F(1,303)=9.789, p=.002) but also an interaction between tasks and belief format (F(1,303)=13.16, p<.001). Here, separate tests indicate that prior biases in the quiz are not affected by format (PriorBias_Mean=-0.18, PriorBias_Distrib=-0.14, t(303)=-0.193, p=.847) whereas prior biases in the perceptual task are different between the two format conditions (PriorBias_Mean=0.01, PriorBias_Distrib=-0.72, t(303)=4.914, p<.001). The observation that participants have similar prior biases between the Mean and Distribution conditions in the quiz task is important, as it will facilitate the comparison of these two conditions regarding how participants will update their beliefs.

Biases in posterior beliefs also exhibit an interaction between belief format and tasks (F(1,303)=14.326, p<.001): in the perceptual task, participants underestimate their scores more in the Distribution format than in the Mean format (PostBias_Mean=0.06, PostBias_Distrib=-0.22, t(303)=2.992, p=.003), whereas in the quiz task the opposite effect of format is found (although not significant, PostBias_Mean=-0.16, PostBias_Distrib=-0.01, t(303)=-1.490, p=.137).

Belief format	**Type of task**	Score (1)	Mean prior belief (2)	Prior bias (3)	Mean posterior belief (4)	Posterior bias (5)
Distribution	Quiz	5.39 (0.11)	5.25 (0.11)	-0.14 (0.10)	5.39 (0.11)	-0.01 (0.06)
Mean	Quiz	5.87 (0.11)	5.69 (0.10)	-0.18 (0.10)	5.71 (0.11)	-0.16 (0.07)
Distribution	Perceptual	7.44 (0.08)	6.72 (0.08)	-0.72 (0.10)	7.22 (0.08)	-0.22 (0.06)
Mean	Perceptual	7.42 (0.08)	7.42 (0.07)	0.01 (0.09)	7.47 (0.07)	0.06 (0.06)

Table 4. Mean of scores, mean prior belief, prior bias, mean posterior belief and posterior bias across participants and variability conditions by type of task and belief format. Standard deviations are reported in parenthesis.

3.2.2. Updating beliefs about mean performance

Since we cannot construct a Bayesian benchmark in the *Mean* condition, we instead rely on the mean difference between the prior and posterior to compare the *Mean* and *Distribution* conditions. In both conditions, we can evaluate how this mean belief difference also depends on good vs. bad news. In particular, we shall test the hypothesis that in comparison to the *Distribution* condition, the *Mean* condition might induce more conservative updating or more asymmetric updating of beliefs (i.e. a greater reaction to good news than to bad news). Figure 6 shows this mean belief difference in the *Mean* and *Distribution* conditions, separately for the quiz task and the perceptual task.

In the quiz task, contrary to the hypothesis formulated above, reactions to good news are not more pronounced but less pronounced in the *Mean* condition, while reactions to bad news are similar between the two conditions.[5] Note that in the quiz task the comparison between the *Mean* and *Distribution* conditions is valid because as shown before, these two conditions show no difference in terms of prior bias here. This is important because it means that the intensity of the feedback received by participants is similar between the two conditions, such that according to Bayes we should expect similar mean updates in the *Mean* and *Distribution* conditions. Besides, since participants' prior beliefs are well-calibrated in the quiz task (PriorBias_Mean=-0.18, t(152)=-1.588, p=.114, PriorBias_Distrib=-0.14, t(151)=-1.279, p=.203), by same reasoning we should expect similar mean updates following good and bad news in theory.

In the perceptual task, by contrast, participants are underestimating their performance overall. Because of this bias, on average good news calls for larger updates than bad news in this task. Moreover, as this underestimation is higher in the *Distribution* condition than in the *Mean* condition, a larger difference between updates for good and bad news is expected in the *Distribution* condition. This expected pattern indeed is confirmed by Figure 6. However, because of this difference in prior biases between the two conditions, testing the hypothesis that belief updating is less biased in the *Distribution* condition than in the *Mean* condition becomes difficult for the perceptual task.

[5] More evidence along the same lines are shown in Supplementary Material.

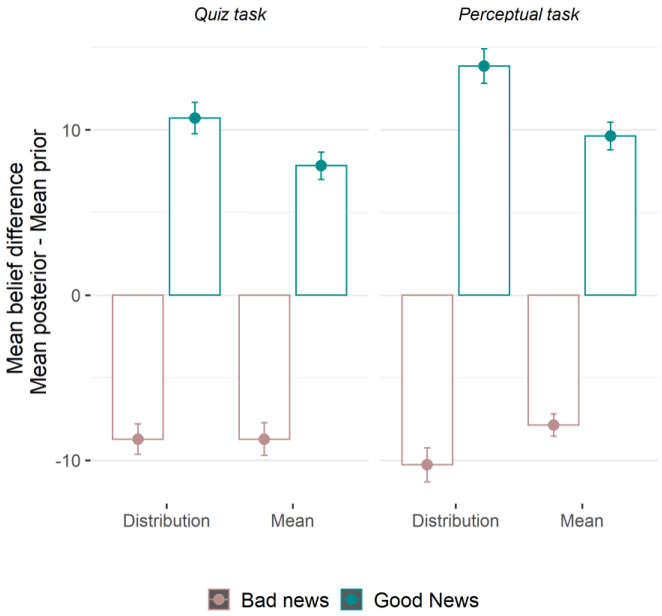

Figure 6. Mean actual updating (mean posterior beliefs - mean prior beliefs) depending on whether the feedback is good news or bad news by type of task and belief format.

3.3. Beliefs about luck

In their paper, G&O provide an indirect measure of luck, that is the difference between feedback and mean posterior. They find that participants' posterior beliefs are higher than the feedback they receive, and conclude that participants should believe that their feedback was unlucky. Here, we construct a more direct measure of luck, by asking participants to report the likelihood of each possible source for the feedback they receive.

Figure 7A illustrates these beliefs about the possible sources for the feedback (Deflated, Wise and Inflated Bob). Participants overall understood the question that was asked as well as the structure of the feedback: Inflated Bob was deemed more likely than Deflated Bob after good news, and less likely after bad news. Moreover, when they receive a feedback equal to their mean prior ("Neutral news") they report that Wise Bob is more likely. Note that Figure 7A shows no sign of the good news-bad news effect in beliefs about luck which should be reflected in bad news being more often associated with bad luck than good news with good luck. On average, participants consider themselves just as lucky when they receive good news as they consider themselves unlucky when they receive bad news.

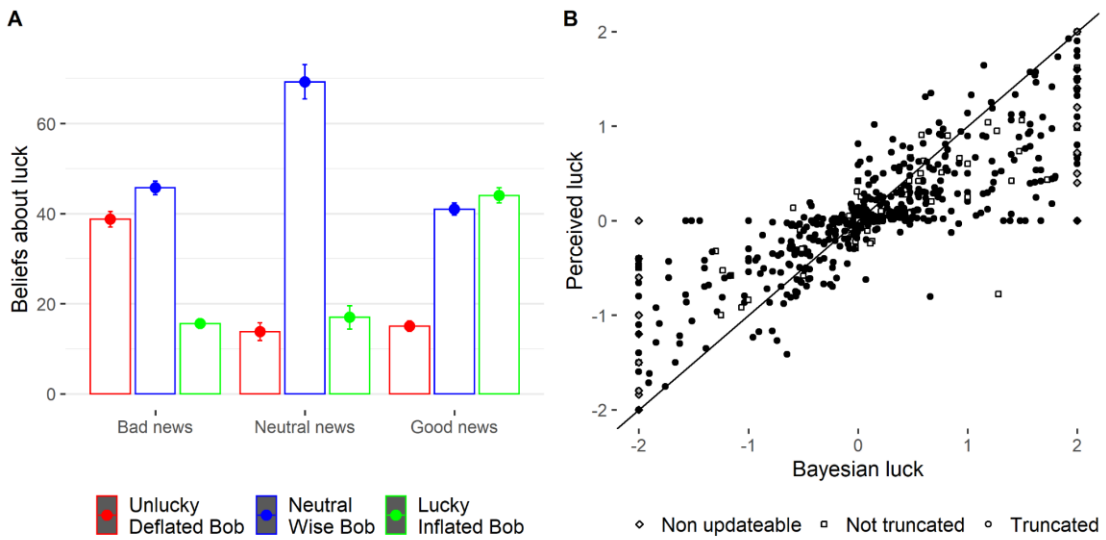

Figure 7. A. Aggregated distribution of belief about the source of the feedback (Deflated Bob in red, Wise Bob in blue and Inflated Bob in green) across experimental conditions depending on whether the feedback is bad news, neutral news, or good news. B. Perceived luck versus Bayesian luck, within condition Distribution. Each participant is represented by 4 observations (1 for each task and 1 for each variability condition).

We now take a more quantitative approach to the analysis of biases in beliefs. Recall that to measure perceived luck, in the *Distribution* condition we evaluate the expected value of the noise in the feedback, given the participant's beliefs about luck and the Bayesian benchmark (see Methods). First, as a sanity check, we note that our measure of perceived luck is correlated with the measure that G&O use in their paper (Cor=.64, p<.001). Then, to evaluate conservatism and asymmetry in perceived luck, we compare our measure to Bayesian luck, that is the perceived luck that would be expressed by a Bayesian agent, via the regression model in eq. (2).[6]

$$Perceived\ luck = a + \beta \times Bayesian\ luck + c \times GN + \gamma \times Bayesian\ luck \times GN \quad (2)$$

Figure 7B illustrates this regression, and reveals some conservatism but no asymmetry in the updating of beliefs about luck. In Table 5, examination of the pooled regression coefficients for conservatism, and asymmetry confirm this statistically across all observations. These results hold when only considering observations in the Truncated category. Furthermore, this result is largely consistent across our 8 experimental conditions.

[6] Note that this regression model is analogous to the one used in eq (1) to evaluate updating of beliefs about performance. Indeed, since participants' prior for perceived luck is necessarily 0, perceived luck is both a posterior belief and an update from the prior.

Conditions	Variability	Pooled	High	Low	High	Low	High	Low	High	Low
	Task	Pooled	Quiz	Quiz	Quiz	Quiz	Percept	Percept	Percept	Percept
	Scoring	Pooled (1)	BDM (2)	BDM (3)	QSR (4)	QSR (5)	BDM (6)	BDM (7)	QSR (8)	QSR (9)
All observations	β	0.57*** (0.05)	0.41*** (0.08)	0.51*** (0.10)	0.78** (0.07)	0.66*** (0.10)	0.47*** (0.10)	0.70** (0.10)	0.54*** (0.09)	0.61** (0.13)
	γ	-0.11 (0.06)	0.12 (0.13)	0.02 (0.13)	-0.53*** (0.12)	-0.23 (0.13)	-0.10 (0.14)	-0.33* (0.13)	-0.10 (0.12)	0.02 (0.15)
	c	0.11 (0.04)	0.19 (0.14)	0.10 (0.15)	0.09 (0.11)	0.08 (0.15)	0.23 (0.16)	0.06 (0.13)	0.07 (0.13)	0.05 (0.18)
	a	-0.04 (0.03)	-0.10 (0.10)	-0.08 (0.10)	0.07 (0.07)	-0.03 (0.10)	-0.17 (0.11)	0.07 (0.11)	0.02 (0.09)	-0.05 (0.16)
	R^2	.69	.65	.70	.76	.67	.64	.70	.71	.75
	N	589	75	76	73	72	74	74	73	72
Only truncations	β	0.57*** (0.04)	0.44*** (0.11)	0.69* (0.12)	0.72*** (0.08)	0.60** (0.14)	0.45*** (0.11)	0.57*** (0.12)	0.54*** (0.08)	0.76 (0.18)
	γ	-0.11 (0.05)	0.16 (0.17)	-0.17 (0.16)	-0.41*** (0.11)	-0.21 (0.18)	-0.06 (0.16)	-0.12 (0.15)	-0.11 (0.12)	-0.18 (0.20)
	c	0.10 (0.05)	0.16 (0.17)	0.02 (0.15)	0.10 (0.10)	0.11 (0.19)	0.22 (0.17)	0.06 (0.14)	0.07 (0.13)	-0.04 (0.20)
	a	-0.03 (0.04)	-0.13 (0.11)	0.03 (0.11)	0.06 (0.06)	-0.05 (0.14)	-0.18 (0.11)	0.02 (0.11)	0.02 (0.09)	0.07 (0.19)
	R^2	.68	.67	.71	.76	.57	.62	.64	.74	.71
	N	484	63	58	62	55	59	62	67	58

Table 5. Regressions examining the conservatism and symmetry of perceived luck by experimental condition, for all observations and for the subsample of Truncated observations only. Participants for whom feedback = mean prior belief are not included in this analysis, as it is considered neither good nor bad news. Asterisks applied to β denote significant difference from one. Asterisks on all other coefficients denote significant difference from zero. Standard errors in parentheses.

4. Discussion

In the present study, we investigate how participants update their beliefs about their own performance in a quiz and in a perceptual task. Our goal is to quantify potential biases of conservatism (under-reaction to feedback in general) and asymmetry (under-reaction more pronounced following bad news). Across various experimental conditions varying the uncertainty of these beliefs, the format in which beliefs are reported (full distribution over scores vs. mean expected performance), the elicitation

rule (BDM vs. QSR), and the task (Quiz vs. Perceptual), we find no consistent evidence for conservatism or asymmetry in belief updating, in the vast majority of participants. Overall, our results thus replicate and generalize the prior findings of G&O. In addition, we introduce a new empirical measure of how participants interpret the feedback they receive as being the result of good or bad luck. This perceived luck shows conservatism but again no asymmetry in our data.

Before we discuss these results in relation to the broader literature on motivated beliefs and optimism in our last paragraph, we wish to return to four methodological aspects that are relevant when comparing our results to G&O.

First, we note that participants in our study exhibit a different bias in the estimation of their scores compared to G&O. More precisely, although our study used similar questions, and although we obtain comparable performance levels, participants' prior beliefs are better calibrated in our data. Participants in our study underestimated slightly their number of correct answers by .16 correct answers while in G&O, participants overestimated their scores in the quiz by 1.14 correct answers. In other words, participants are overconfident in G&O but not in our study. Whereas the reasons for this difference are unclear, a number of contextual factors differ between the two studies: our study was conducted on-line on isolated subjects, during the Covid era, whereas participants in G&O were taking part in group sessions in the laboratory, at a time where Barack Obama was president. Importantly, as this bias affects the relative frequency of good news and bad news, the fact that we replicate the original results of G&O in a different setting with different biases is reassuring regarding the robustness of these results.

Second, one original methodological aspect in G&O is the sorting of observations in three categories (*Not updateable*, *Not truncated* and *Truncated*) with respect to whether prior and posterior beliefs are compatible with the feedback. In both our study and G&O, most observations fall in the *Truncated* category, and importantly observations in that category conform to Bayesian updating.[7] We also note that we do not replicate the updating biases found by G&O in the *Not updateable* category, but this category contains very few observations in their study. A few observations fall in the *Not Truncated* category and seem to show some conservatism in our data. Since this category corresponds to cases where parts of the posterior beliefs are outside of the range of scores that are compatible with the feedback, our interpretation is that these observations likely correspond to errors of inattention or to participants misunderstanding the nature of the feedback. In our data, these observations do not seem to be randomly distributed across participants. Indeed, whereas only 11% of observations fall in this category, participants who have one *Not Truncated* observation in a given condition have on average 49% of *Not Truncated* observations in the remaining 3 conditions.

Third, although we find no effect of ego-relevance on belief updating like G&O, we must acknowledge that our empirical measure of ego-relevance differs from what is typically used in the literature. Most studies rely on beliefs about oneself in the ego-relevant case and on beliefs about another participant or a random device in the non-ego-relevant case. Instead, our study focuses on beliefs about performance in two tasks, we ask participants to indicate whether these tasks are relevant to them, and to evaluate the influence of ego-relevance we use participants for whom one task is deemed as ego-relevant and the other task is not. However, we note that in our data, participants tend to evaluate the quiz task as more relevant when they perform well in this task, which is consistent with Drobner (2022) who shows that individuals manipulate their belief about the ego-relevance of a given event depending on the feedback they receive. Thus, the definition of ego-relevance is not independent from performance in our case, which might confound our results. To strengthen our claim, we conducted a supplementary study where we collected more data by replicating the condition "Other" in G&O: instead of rating their own performance, participants were asked their beliefs about the performance of a randomly selected other participant. Much like G&O, we find that participants in this condition are slightly

[7] To be clear, focusing on the *Truncated* category only does not enforce Bayesian updating, as conservatism and asymmetry could still be found in this category, in theory.

oversensitive feedback (significantly so when only the Truncated observations are included) and overall, update symmetrically good and bad news. We further compare this supplementary study to our main study and find no significant interaction between this condition and biased belief updating, confirming that participants update similarly ego relevant and non-ego relevant performance beliefs (see Supplementary Materials 6 for more details about the design and the analyses of this supplementary study).

Fourth, one original contribution of our study is to develop a measure of perceived luck regarding the feedback received. Unlike beliefs about performance, this measure of perceived luck does exhibit conservatism (but no asymmetry) in our data. It is unclear why this measure of perceived luck behaves differently from beliefs on performance. Yet, we may speculate that judging the likelihood of the possible sources for the feedback is more complex, perhaps less ecological also, than evaluating one's own performance. One other possibility is that the objective and explicit prior regarding the distribution of the feedback given by the experimenter is also less uncertain than the subjective prior that is formed by the participant in the case of beliefs about performance. These different factors should in theory contribute to a greater anchoring effect for the measure of perceived luck.

To conclude this discussion, we shall take a broader perspective and discuss our results with respect to the literature on motivated beliefs. This term generally refers to the idea that agents might believe desirable outcomes to be more likely, because such beliefs bring utility to the agent (either directly through self-image, as in Köszegi, 2006, or anticipatory utility, as in Brunnermeier & Parker, 2005; or indirectly via motivation, as in Benabou and Tirole 2002, Compte and Postlewaite, 2004; or signals sent to others, as in Schwardmann & van der Weele, 2019). The asymmetric updating phenomenon under scrutiny here is suggested to be a consequence of motivated beliefs, along with other biases such as unrealistic optimism (Möbius et al., 2022; Sharot et al., 2011; Weinstein & Klein, 1996) or desirability bias (Kunda, 1990), depending on the inferential process considered. However, the strength of the empirical evidence supporting the idea of motivated beliefs remains debated (Hahn & Harris, 2014), with some critics raising methodological concerns e.g. the use of extremely unlikely events (Harris & Hahn, 2011), and some experiments not replicating the observation of motivated beliefs (Barron, 2021; Coutts, 2019; Ertac, 2011; Grossman & Owens, 2012; Harris et al., 2017; Shah et al., 2016). In addition, alternative rational explanations have been proposed to explain belief biases (Hahn & Harris, 2014). In this literature, the originality of G&O is to evaluate motivated beliefs for judgments on absolute performance. They report no evidence for motivated beliefs, and the present study extends these results while varying the uncertainty in the participants' priors, the tasks to perform, the beliefs' elicitation rules and by proposing a direct measure of the feedback credibility (beliefs about luck) which allows to test in another way asymmetric updating. The absence of any asymmetric updating in all our conditions provides strong evidence against the existence of motivated beliefs in the context of individual judgments about one's performance.

In a recent attempt to understand the heterogeneous results observed in the asymmetric updating literature, Barron (2022) concluded that neither differences in information structure, priors, domains and stake sizes among the different studies in this literature could fully account for those mixed results. Although a detailed investigation of the contextual factors that may foster asymmetric updating is outside of the scope of the present paper, our work actually points to another possible explanation. More precisely, it might be that motivated belief updating (and consequently, overweighting of good news compared to bad news) is more likely to be triggered in contexts involving social comparisons between individuals, where the beliefs to be updated are about one's performance relative to other individuals. This hypothesis would be in line with two recent working papers which evaluate updating on absolute (Coffman et al, 2019) and relative (Coffman et al, 2021) performance beliefs. Although the authors do not specifically study motivated belief updating, in Coffman et al. (2019), there is no evidence that participants overweight good compared to bad news (results rather tend towards a reverse asymmetry) while in Coffman et al (2021), there is some evidence of such overweighting of good news.

In order to properly test this hypothesis in our setup, we conducted another supplementary study where instead of rating their performance in absolute terms, participants were asked about their rank relative to a set of randomly selected other participants, as in Coffman et al. (2021). However, it turns out that we do not find any evidence for asymmetric updating in this case either and that updating is not different from updating in our main study (see Supplementary Materials 7 for more details about the design and the analyses of this supplementary condition). Although our hypothesis is unlikely to solve the big puzzle of the asymmetric updating literature, a more in-depth comparison between absolute and relative performance belief updating would be needed to dismiss its relevance more firmly. In particular, one would need to test this hypothesis with a more standard (and perhaps simpler) binary signal structure, used in most studies which do find evidence of asymmetric belief updating (Eil & Rao, 2011; Charness & Dave, 2017; Coffman et al, 2021; Mobius, 2022).

References

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

Becker, G. M., & DeGroot, M. H. (1974). Measuring Utility by a Single-Response Sequential Method (1964). In *Economic Information, Decision, and Prediction : Selected Essays : Volume I Part I Economics of Decision* (p. 317–328). Springer Netherlands.

Benabou, & Tirole. (2002). Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics*, *117*(3), 871- 915.

Benoît, J.-P., & Dubra, J. (2011). Apparent Overconfidence. *Econometrica*, *79*(5), 1591- 1625. https://doi.org/10.3982/ECTA8583

Brunnermeier, & Parker. (2005). Optimal expectations. *American Economic Review*, *95*(4), 1092- 1118.

Burnham, & Anderson. (2002). *Model Selection and Multimodel Inference—A Practical Information-Theoretic Approach*. Springer, New York, NY.

Buser, T., Gerhards, L., & van der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, *56*(2), 165- 192. https://doi.org/10.1007/s11166-018-9277-3

Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological cybernetics*, *107*(6), 711–719.

Charness, G., Dave, C., 2017. Confirmation bias with motivated beliefs. *Games and*

Economic Behavior, 104, 1–23.

Coffman, Katherine B., Manuela Collis, and Leena Kulkarni. (2019) Stereotypes and Belief Updating. *Harvard Business School Working Paper*, N° 19-068

Coffman, Katherine B., Paola U. Araya, and Basit Zafar. (2021) A (Dynamic) Investigation of Stereotypes, Belief-Updating, and Behavior. *NBER Research paper,* N° 29382

Compte, & Postlewaite. (2004). Confidence-Enhanced Performance. *American Economic Review*, *94*(5), 1536- 1557.

Coutts, A. (2019). Good news and bad news are still news : Experimental evidence on belief updating. *Experimental Economics*, *22*(2), 369- 395. https://doi.org/10.1007/s10683-018-9572-5

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*(4), 568–584.

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

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz & Symposium on Cognition. Annual (Éds.), *Formal representation of human judgment*. Wiley.

Eil, & Rao. (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. https://doi.org/10.1016/j.jebo.2011.05.01

Fischhoff, & Beyth-Marom. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*.

Grieco, D., & Hogarth, R. M. (2009). Overconfidence in absolute and relative performance :

The regression hypothesis and Bayesian updating. *Journal of Economic Psychology*, *30*(5), 756- 771.

Grossman, Z., & Owens, D. (2012). An unlucky feeling : Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, *84*(2), 510- 524. https://doi.org/10.1016/j.jebo.2012.08.006

Hahn, & Harris. (2014). What does it mean to be biased : Motivated reasoning and rationality. In *Psychology of learning and motivation* (Vol. 61, p. 41- 102).

Harris, de Molière, Soh, & Hahn. (2017). Unrealistic comparative optimism : An unsuccessful search for evidence of a genuinely motivational bias. *Plos One*, *12*(3).

Harris, & Hahn. (2011). Unrealistic optimism about future life events : A cautionary note. *Psychological review*, *118*(1), 135.

Harrison, Martínez-Correa, Swarthout, & Ulm. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization*, *134*, 430- 448.

Hoffrage. (2017). Overconfidence. In *Cognitive illusions : Intriguing phenomena in thinking, judgment and memory* (p. 291–314). R. F. Pohl.

Huck, & Weizsäcker. (2002). Do players correctly estimate what others do? : Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization*, *47*(1), 71- 85.

Köszegi. (2006). Ego utility, overconfidence and task choice. *Journal of the European Economic Association*, *4*, 673- 707.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it : How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121- 1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kunda. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Lichtenstein, & Slovic. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(1), 46- 55.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and*

Social Psychology, 37(11), 2098–2109.

Menkhoff, Schmeling, & Schmidt. (2013). Overconfidence, experience, and professionalism : An experimental study. *Journal of Economic Behavior & Organization*, *86*, 92- 101.

Möbius, M., Niederle, M., Niehaus, P., & Rosenblat, T. (2022). Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 387- 392.

Peterson, & Miller. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, *70*(1), 117- 121.

Qu. (2012). A mechanism for eliciting a probability distribution. *Economics Letters*, *115*(3), 399- 400.

Scheier, M. F., ZZZ, & ZZD. (1989). Dispositional optimism and recovery from coronary artery bypass surgery : The beneficial effects on physical and psychological wellbeing. *Journal of Personality and Social Psychology*, *57*, 1024–1040.

Schwardmann, P., & van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, *3*(10), 1055- 1061. https://doi.org/10.1038/s41562-019-0666-7

Shah, Harris, & Hahn. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, *90*, 71- 127.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 379- 423.

Sharot, Korn, & Dolan. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), 941–945.

Strunk, Lopez, & DeRubeis. (2006). Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behaviour Research and Therapy*, *44*(6), 861- 882.

Svenson, O. (1981). Are We All Less Risky and More Skillful than our Fellow Drivers? *Acta Psychologica*, *47*(2), 143- 148.

Taylor, Kemeny, Reed, Bower, & Gruenewald. (2000). Psychological resources, positive illusions, and health. *American Psychologist*, *55*(1), 99–109.

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

Weinstein, & Klein. (1996). Unrealistic optimism : Present and future. *Journal of Social and Clinical Psychology*, *15*(1), 1- 8.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence : Individual differences in performance estimation bias. *Psychonomic Bulletin & Review*, *4*(3), 387- 392.