Title

Mental Models and Learning: The Case of Base-Rate Neglect

Permalink

https://escholarship.org/uc/item/8cb387t8

Journal

American Economic Review, 114(3)

ISSN

0002-8282

Authors

Esponda, Ignacio
Vespa, Emanuel
Yuksel, Sevgi

Publication Date

2024-03-01

DOI

10.1257/aer.20201004

Mental Models and Learning: The Case of Base-Rate Neglect

Ignacio Esponda Emanuel Vespa Sevgi Yuksel*

December 5, 2023

Abstract

Are systematic biases in decision making self-corrected in the long run when agents are accumulating feedback informative of optimal behavior? This paper focuses on a canonical updating problem where the dominant deviation from optimal behavior is base-rate neglect. Using a laboratory experiment, we document persistence of suboptimal behavior in the presence of feedback. Using diagnostic treatments, we study the mechanisms hindering learning from feedback. We investigate the generalizability of these results to other settings by also studying long-run behavior in a voting problem where failure to condition on being pivotal generates suboptimal behavior. Our findings provide insights on what types of mistakes should be expected to be persistent in the presence of feedback. Our results suggest mistakes are more likely to be persistent when they are driven by incorrect mental models that miss or misrepresent important aspects of the environment. Such models induce confidence in initial answers, limiting engagement with and learning from feedback. These results have implications for how policies should be designed to counteract behavioral biases.

*Esponda: Department of Economics, UC Santa Barbara (email: iesponda@ucsb.edu). Vespa: Department of Economics, UC San Diego (email: evespa@ucsd.edu). Yuksel: Department of Economics, US Santa Barbara (email: sevgi.yuksel@ucsb.edu). We would like to thank the editors, anonymous referees and many seminar participants for helpful comments. In addition, we would like to thank Guillaume Fréchette, Muriel Niederle, and Ryan Oprea for numerous conversations in the process of writing this paper. Invaluable research assistance was provided by Dingyue Liu and Caroline Zhang. We are grateful for financial support from the UCSB Academic Senate. The experimental data collected for this paper had protocols approved by IRB at UC Santa Barbara (protocol number 3-19-0900) and UC San Diego (protocol number 801766).

1 Introduction

Behavioral economics has accumulated a wealth of evidence documenting systematic biases in decision making. An important question is whether such biases are self-corrected in the presence of feedback. On the one hand, biases might vanish with experience if agents are accumulating evidence informative of optimal behavior. On the other hand, this type of learning presumes agents are attentive to the feedback they are experiencing, willing and able to adjust their behavior in response to it. A growing empirical and theoretical literature challenges this position by emphasizing how initial misconceptions can have long-lasting effects on how people learn from their experiences.[1]

In this paper, we present results from a laboratory experiment designed to study optimality of long-run behavior in the presence of feedback and bring to light the different mechanisms that hinder learning from feedback. The experiment has two crucial features. First, we consider a baseline treatment where subjects face a decision problem where they are given information that would enable them to solve the problem optimally. However, such information is known to be used incorrectly and produces biased behavior. We study the evolution of this bias when subjects face multiple rounds and receive transparent feedback. Second, we compare behavior in this treatment to a control treatment in which information inducing biased behavior is withheld from the subjects. In the absence of such information, subjects cannot initially solve the problem, but can use feedback to learn about optimal behavior. This design allows us to study the extent to which initial misconceptions induced by payoff-relevant information about the problem can inhibit learning from feedback.

In our baseline treatment, information we provide to the subjects induces one of the most well-documented biases in the literature, base-rate neglect. As a motivating example (adapted from Kahneman & Tversky 1972), consider a person who is tested for a disease. The disease has a prevalence of 15 percent in the general population and the test has an accuracy of 80 percent.[2] With these primitives, the chance that the person is sick conditional on a positive test result is 41 percent, but the literature has repeatedly documented that many subjects (and doctors!) incorrectly consider this chance to be 80 percent (see Benjamin (2019) for a survey). Because such beliefs *completely* fail to take into account the unconditional probability of the disease, we refer to this bias as *perfect* base-rate neglect (pBRN).

While BRN is not the only deviation from the Bayesian benchmark observed in the data, it is

[1] For recent theoretical and empirical contributions see Esponda & Pouzo (2016) and Hanna, Mullainathan & Schwartzstein (2014), respectively. For more references, see discussion of the literature.

[2] The probability of a positive test result conditional on the person being sick (not sick) is 80 (20) percent.

the overwhelmingly dominant one: More than half our subjects' initial beliefs are consistent with pBRN. The experimental design involves subjects facing the same decision problem for 200 rounds. In each round, a new state is randomly selected and a signal is drawn. Subjects submit beliefs conditional on the signal, and observe the true state at the end of the round. The interface also displays a record of all past outcomes. In our baseline treatment, labeled as *Primitives*, subjects are presented with the above problem (albeit with a more neutral framing) and informed of the primitives (i.e., the 15 percent prior and the 80 percent accuracy of the signal) so that, in principle, they could provide the correct response of 41 percent (conditional on a positive signal) from the very first round.

Our focus is on the optimality of long-run behavior in response to feedback, specifically how close beliefs are to the Bayesian benchmark after 200 rounds. We find that, at the aggregate level, the adjustment is slow and partial. For example, the average belief conditional on a positive signal, which starts at 64 percent in round one, drops to 54 percent by round 200. While the adjustment is significant, it also remains substantially above the Bayesian benchmark of 41 percent, implying that the wrong state is persistently judged to be more likely. These results show that subjects' incorrect understanding of how to make use of the primitives have long lasting effects even in a context where there is abundant evidence (feedback about past outcomes in this case) that is informative about optimal behavior.

However, it is difficult to interpret long-run beliefs in *Primitives* on its own. We need a benchmark that captures how much subjects could have learned from the feedback provided in 200 rounds in the absence of any other information that might induce an incorrect understanding and hence bias behavior. In other words, we need a counterfactual environment where subjects need to rely on feedback alone to determine optimal behavior. With this aim, we conduct a control treatment, labeled as *NoPrimitives*, in which subjects face the same updating task described in the *Primitives*, except that they are not provided with the primitives. That is, subjects receive the same description of the task but are not given the specific values for the prior and the accuracy of the signal. As in the baseline treatment, we let subjects experience the realization of the state and the signal in every round for a total of 200 rounds. The feedback subjects receive is structurally the same in both treatments because it is generated by the same primitives, and it is exogenous to the subjects' beliefs.

We find an important treatment effect after 200 rounds with respect to the accuracy of beliefs: In aggregate, beliefs in the control treatment (*NoPrimitives*) are closer to the Bayesian benchmark relative to beliefs in the baseline treatment (*Primitives*). For example, the average belief conditional on a positive signal is at 46 percent in *NoPrimitives* which is eight percentage points lower than the value in *Primitives*.[3] Moreover, the treatment effect disappears if we exclude subjects who provide the pBRN answer in the initial round, suggesting that, of all initial misconceptions induced by the *Primitives* treatment, it is principally those inducing the pBRN beliefs in round one that hinder learning from feedback.

We then turn to understanding the channels through which learning from feedback is made more difficult in *Primitives*. We conduct additional treatments and make use of a learning model to provide insights on mechanisms.

First, we investigate whether initial misconceptions, induced by information on the primitives, hinder learning from feedback by endowing subjects in *Primitives* with unjustified high confidence in their initial responses. To test this, we run a diagnostic treatment that is identical to *Primitives* except for one small difference. At the end of round one, we tell subjects (to whom the message applies) that their initial responses are not correct. Otherwise, subjects experience 200 rounds of feedback in the same way. The message has a large impact on how close beliefs are to the Bayesian benchmark after 200 rounds of feedback. The average belief conditional on a positive signal drops to 43 percent, 11 percentage points lower than the value in *Primitives*. In fact, all subjects, including those with initial pBRN beliefs, are capable and willing to learn from feedback. Together with our earlier findings, this suggests that subjects with high confidence in their initial pBRN beliefs play a critical role in inhibiting learning from feedback in *Primitives*.

Second, we ask whether initial misconceptions, induced by information on the primitives, hinder learning from feedback by also reducing subjects' attentiveness to the feedback available to them.[4] Specifically, we conduct a set of diagnostic treatments to examine whether information on the primitives impacts engagement with feedback. These treatments are identical to *Primitives* and *NoPrimitives*, except that we allow subjects to "lock in" their responses at any point during the 200 rounds. Once responses are locked-in, they are automatically implemented for all future rounds. This lock-in decision gives us a simple measure of engagement by revealing how many rounds of feedback subjects are willing to see. Our results highlight large differences in engagement with feedback. When provided with the primitives, only half the subjects choose to see more than 20

[3] The finding that long-run behavior is approximately optimal in *NoPrimitives* is in line with the frequentist hypothesis in evolutionary psychology (Cosmides & Tooby 1996), which states that some reasoning mechanisms in humans are naturally designed to use frequency information. It is also consistent with studies establishing that animal foraging behavior is approximately optimal despite the primitives of the environment being unknown, a finding sometimes attributed to the ability to track frequencies (e.g., Lima (1984)).

[4] Feedback in these treatments is presented on a round-by-round basis. The design also provides subjects with a record of all past outcomes. By attentiveness, we mean going beyond merely observing outcomes, but also aggregating them in a manner that may allow the agent to learn from them.

rounds of feedback and only four percent choose to observe all 200 rounds. By contrast, without information on the primitives, 94 percent of subjects choose to see more than 20 rounds of feedback and 31 percent of subjects see all 200 rounds.

Third, the impact of initial confidence and the decision to engage with data crucially depends on the cost of learning, and so we investigate the extent to which these costs hinder learning. We run two more treatments, which are identical to *Primitives* and *NoPrimitives* except that we provide feedback on a round-by-round basis in an aggregated and processed way. Specifically, in each round, we summarize feedback observed up to that point in an easy-to-read table; in addition, we report the empirical frequency of the state conditional on each signal. These treatments reveal how behavior evolves differently with and without information on primitives when the cost of processing feedback is effectively lowered to zero. Results show that when feedback is presented in this way, subjects are able to learn more in both treatments. Average beliefs conditional on a positive signal drop to 44 and 41 percent, in the treatments with and without information on the primitives, respectively. Then, by making use of a simple learning model and combining results from the new treatments with the earlier ones, we separately identify the degree to which our earlier results on the long-run differences between *Primitives* and *NoPrimitives* are due to (i) higher confidence in initial response; and (ii) lower attentiveness to feedback in the former environment. Our results suggest that both channels play an equally important role.

Finally, we study whether subjects in *Primitives* who respond to feedback, simply adjust their beliefs to be consistent with observed frequencies, or whether they gain a deeper understanding of why their initial answers were wrong. We do so by including one last updating problem where the prior and the accuracy of the test are changed, and subjects in both *Primitives* and *NoPrimitives* are equally informed about the new primitives. We find that the treatment effect reverses: average beliefs in *Primitives* are closer to the Bayesian benchmark than in *NoPrimitives.* While learning is partially transferable to this new setting, a non-negligible amount of base-rate neglect remains in *Primitives*, though a much higher proportion appears in *NoPrimitives.*

Throughout the paper, we use the term 'misconception', or alternatively incorrect 'mental model', broadly to refer to an agent's incorrect initial understanding of the environment that misses or misrepresents important aspects of reality while endowing the agent with confidence in their initial answer.[5] We find persistent failures to learn in information-rich environments and that these failures are driven by confidence in an incorrect initial answer. Confidence hinders learning

[5] In a general sense, different types of initial misconceptions can arise in any setting, with or without information on primitives. But, by contrasting such treatments (with and without information on the primitives), we are able to study the long-run implications of misconceptions that manifest in one setting but not the other.

both by making subjects less responsive (put less weight) on new information and by lowering attentiveness to such information.

These findings provide insights on what other types of mistakes might fail to be self-corrected with experience. Our results suggest that mistakes that are driven by an incorrect understanding of the environment that misses or misrepresents some aspects of reality might not be corrected. On the other hand, not all mistakes are driven by incorrect mental models, such as those that arise because it is cognitively costly to identify optimal behavior. In such cases, our findings suggest that the agent will be self-aware of the possibility of a mistake, and will be more open to engaging with feedback and correcting their behavior.

We conclude by assessing the generalizability of our results and testing our hypothesis about the types mistakes that are likely to persist in a new environment. We conduct four more treatments in a setting involving a voting decision where an agent, by conditioning on the case when her vote is pivotal, could identify that there is a dominant action. However, the framing of the problem is such that an agent who fails to condition on this contingency (pivotality) would incorrectly perceive the decision as reflecting risk preferences.[6] As in our original treatments, we elicit initial and long-run responses in the presence of feedback. First, replicating our main result in a new setting, we document higher rates of optimal behavior in the long-run in a treatment where subjects were not given the primitives relative to one where they were. This result reaffirms the main message of the paper that mistakes that are driven by incorrect understanding of the environment that miss or misrepresent some aspects of reality are difficult to correct. In our last two treatments, we present the same voting problem but with the options deliberately described in a more complicated manner. This makes the initial misconception (that the problem represents a choice on risk) less apparent. According to our hypothesis about the types of mistakes that are more likely to persist, the complex description should make it more likely that subjects are aware of the possibility of a mistake in their initial responses, and this should in turn improve learning. Consistent with our hypothesis, we find that subjects are less confident in the complex framing, and do equally well in the long run with or without information on the primitives.

Connections to the literature

The themes explored in this paper, in terms of how learning from past experiences is necessarily shaped by our initial understanding of the world, connect with a few different literatures. First, our results provide support for a growing literature in economics that studies the implications of

[6] The setting is based on the problem studied in Ali, Mihm, Siga & Tergiman (2021).

incorrect or misspecified models. A central premise of this literature is that the degree to which an agent learns from past experiences is constrained by her initial misspecified model.[7] There is also a related literature that models why misrepresentations can arise in the first place (e.g., Gennaioli & Shleifer (2010), Bordalo, Gennaioli & Shleifer (2013), and Gabaix (2014)) and emphasizes cognitive difficulties associated with comprehending and integrating important features of the environment to the decision making process.[8] Such cognitive difficulties may explain agents' reliance on simpler (but incorrect) mental models. Furthermore, our result that some agents change their model with feedback but others do not speaks to a small literature that studies how agents question and change their models of the world (e.g., Ortoleva (2012), Montiel Olea et al. (2022), Fudenberg & Lanzani (forthcoming), He & Libgober (2023).)

Second, an emerging literature endogenizes attentiveness to payoff-relevant features of the environment when there are information processing costs. The literature on rational inattention (e.g., Sims 2003; Caplin & Dean 2015) assumes agents have rational expectations about the value of such information, but trade off this value against learning costs. Building on this intuition, but allowing agents to be systematically misguided in how they assess the value of information, Schwartzstein (2014) and more recently Gagnon-Bartsch, Rabin & Schwartzstein (2021) model the learning process of an agent who channels her attention to a subset of events that are deemed relevant by her (potentially incorrect) mental model, blocking out other types of information. Consistent with our experimental results, these theory papers demonstrate how suboptimal behavior can persist in the long run even when there are negligible attention costs because agents have mistaken initial views on what and how they can learn from feedback. Following the language of Handel & Schwartzstein (2018), such failures in learning would not be driven by "frictions" that are associated with costly information processing, but "mental gaps" that are resulting from misjudgments about the value of information.[9]

Even in the absence of direct information-processing costs, there could be other behavioral forces that influence an agent's engagement with feedback. For example, either due to motivated beliefs (e.g. Bénabou & Tirole 2003; Brunnermeier & Parker 2005; Köszegi 2006) or simply due

[7] For recent examples, see Esponda & Pouzo (2016), Fudenberg, Romanyuk & Strack (2017), Bohren & Hauser (2021), and Heidhues, Kőszegi & Strack (2018).

[8] See for example, Eyster & Weizsäcker (2010), Cason & Plott (2014), Esponda & Vespa (2014), Louis (2015), Dal Bó et al. (2018), Ngangoué & Weizsäcker (2021), Esponda & Vespa (2023), Martínez-Marquina, Niederle & Vespa (2019), Araujo, Wang & Wilson (2021), Martin & Muñoz-Rodriguez (2019), Moser (2019), Graeber (2022), Enke & Zimmermann (2019), Enke (2020), Bayona, Brandts & Vives (2020).

[9] While there is limited empirical evidence on this, our paper is not the first to show that agents can be suboptimally inattentive to features of the environment that are payoff relevant. For instance, Hanna et al. (2014) find that Indonesian seaweed farmers persistently fail to optimize along a dimension (pod size) despite substantial evidence because they fail to examine the data in a way that would suggest its importance. See Gagnon-Bartsch, Rabin & Schwartzstein (2021) for more examples.

to a desire for consistency (Falk & Zimmermann 2018), agents might be reluctant to adjust their behavior in response to past outcomes.[10] These different literatures share a common insight that initial misconceptions can inhibit learning by impacting the way agents engage with the data, and our experiment provides strong evidence for this channel.

Our paper also relates to a literature that studies long-run outcomes in the presence of feedback. In many of these cases, it is challenging to identify the mechanisms that hinder learning from feedback. For example, learning in strategic settings is complicated by the fact that agents may also have to make inferences about the strategies of others, and these strategies may change over the course of the experiment. Moreover, in many problems, feedback is often partial, noisy, endogenous to the subject's choices, or subjects may face sample selection issues (e.g., Huck, Jehiel & Rutter 2011, Esponda & Vespa 2018; Enke 2020; Araujo, Wang & Wilson 2021; Barron, Huck & Jehiel 2019). Yet another example of why learning from feedback might be difficult is the case of an agent who makes choices such that the collected information cannot challenge her model of the world (e.g. Dekel, Fudenberg & Levine 2004; Fudenberg & Vespa 2019).[11] To control for these issues, we focus on simple decision problems in which feedback is simple, transparent and exogenous to the subjects' choices.

There is also a large literature on the specific bias that we primarily focus on, base-rate neglect, initiated by Kahneman & Tversky (1972) and recently surveyed in Benjamin (2019), which also summarizes evidence on the pervasiveness of this bias in important settings (e.g., medical diagnosis, court judgments).[12,13] The broader literature largely abstracts from responses to feedback and learning. A small literature in psychology studies base-rate neglect in the presence of feedback, but this literature focuses on the evolution of beliefs when subjects are not given the primitives and only observe outcomes from a natural sampling process. To our knowledge, there has not been an experiment contrasting learning in treatments with and without primitives with the goal of studying the role initial misconceptions play in the persistence of biases.[14]

[10] See Bénabou & Tirole (2016) for an extensive review of this literature. Recently, Zimmermann (2020) and Huffman, Raymond & Shvets (2022) study the connection between persistent overconfidence and distortions in memory through selective recall when there is repeated feedback.

[11] More details on the recent experimental papers studying subjects' response to feedback is included in Online Appendix A.

[12] The public debate on effectiveness of vaccines provides a perfect example of how base-rate neglect can have dire consequences in a high-stakes environment. Major news organizations were reporting data on vaccine effectiveness failing to properly account for base-rate information (e.g. link1). These types of misrepresentations of the data lead to a public effort to train people to correctly account for base-rates (e.g. link2)

[13] There is also a literature related to the voting problem that we study in our last treatments. As a reference, see Esponda and Vespa (2014, 2023), and Ali, Mihm, Siga & Tergiman (2021).

[14] More detailed discussion of the psychology literature studying base-rate neglect in the presence of feedback is included in Online Appendix A.

2 Experimental design

We designed the experiment to serve two main goals. First, the design allows us to study the persistence of a well-documented bias (BRN) in the presence of feedback in a simple framework, where feedback is natural, informative and independent of the subjects' choices. Second, the design includes a control treatment (without primitives) in which feedback is structurally the same, but mistakes resulting from incorrect use of primitives (such as BRN) are not possible. Thus, the control treatment provides us with a benchmark on subjects' long-run beliefs when feedback is the only information provided to them.

In this section, we describe the overarching design framework used in all treatments and the details associated with the first two parts of the core treatments, which test the central hypothesis in the paper. The remaining two parts of the core treatments and nine additional supporting treatments are introduced in subsequent sections and designed to study the mechanisms underlying these results and the generalizability of these results to other settings.[15]

I. Updating task: Round One

This first part, referred to as round one, introduces the main belief-updating task. The task consists of updating beliefs about the chance that a randomly selected project is a success or failure conditional on a signal being positive or negative. There are 100 projects in total, 15 of which are successes and the remaining 85 are failures, implying a prior (ex-ante probability that a randomly selected project is a success) of 15 percent. After randomly drawing a project, the interface produces a signal, positive or negative, with a reliability of 80 percent. This means that if the project is a success (failure), the signal, which is framed as a test result, will be positive (negative) with 80 percent chance and negative (positive) with 20 percent chance. This parameterization (prior $p = .15$, reliability of signal $q = .8$) corresponds to the classic parameterization of Kahneman & Tversky (1972).

The core of our experimental design consists of two between-subject treatments which differ only in the instructions provided in this part. The treatments, referred to as *Primitives* and *NoPrimitives*, vary in whether subjects are provided with the primitives of the problem or not. All other parts of the instructions, in this part and in all subsequent parts, are identical.

In *Primitives*, subjects know that 15 projects are successes and 85 projects are failures and

[15] A full description of the experimental design for all treatments is provided in Online Appendix B. For the full details that allow an exact replication of our experiment, we refer the reader to the Online Procedures Appendix, where we include instructions and screenshots relating to each part.

that the signal has a reliability of 80 percent. In *NoPrimitives*, subjects know that some projects are successes and some are failures, but they are *not* told how many are successes and how many are failures, and they are also not told the reliability of the signal. In both treatments, using the strategy method, we ask subjects to submit two assessments: (1) the belief that the project is a success conditional on the signal being positive (B_{Pos}), and (2) the belief that the project is a success conditional on the signal being negative (B_{Neg}). In this round and in all future belief-elicitation rounds, subjects are incentivized using a standard incentive-compatible mechanism.[16]

In *Primitives*, subjects could in principle use Bayes' rule to provide the correct answer. Given the prior $p = .15$ and the reliability of the signal, $q = .8$, the Bayesian posterior that the project is a success conditional on a positive signal is, in percentage terms, $B^{Bay}_{Pos} = \frac{pq}{pq+(1-p)(1-q)} \times 100\% = 41\%$. Similarly, the Bayesian posterior that the project is a success conditional on a negative signal is $B^{Bay}_{Neg} = 4\%$. The literature, however, finds that many subjects respond by fully ignoring the prior (treating it as uniform), a response that we call perfect Base Rate Neglect (pBRN) and we denote in percentage terms by $(B^{pBRN}_{Pos}, B^{pBRN}_{Neg}) = (80, 20)$. In *NoPrimitives*, there is no correct way to respond and there is of course no way to suffer from BRN, since the primitives are not provided. To avoid confusion, we specifically tell subjects in this treatment that clearly there is not enough information at this point to make an informed decision.

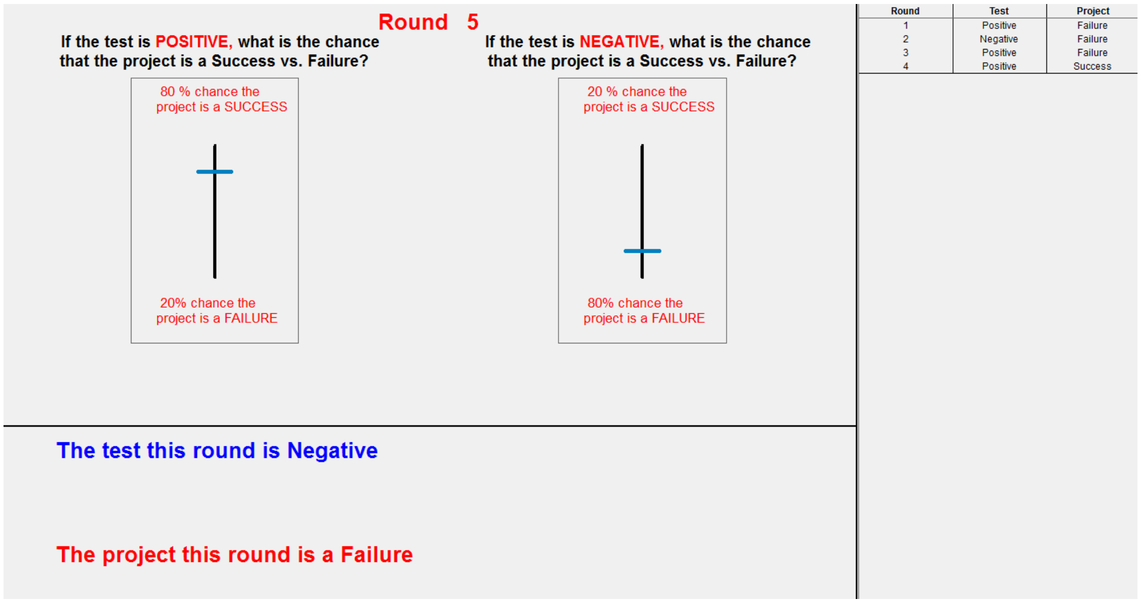
II. Learning: Repetition of updating task, rounds 2-200

This part of the experiment allows us to study how experience and feedback affects beliefs in each treatment. In this part, subjects repeat the task they faced in round one for another 199 rounds.[17] The reliability of the signal and the prior are the same in all rounds and equal to round one ($p = .15, q = .8$), and the state is drawn independently and with replacement in every round.

This part is divided into two phases. The first phase encompasses rounds 2 through 100. At the end of each round, subjects receive feedback on the signal (signal is positive vs. negative) and state (project is a success vs. failure) realizations. The right side of the screen includes a history box that records the signal and state realizations observed in each of the past rounds. Figure 1 shows a screen shot of round 5. In the top-left of the screen, the subject submits a belief conditional on a positive signal and a belief conditional on a negative signal. The figure shows a subject who

[16] Belief elicitation has been combined with the strategy method in a number of prior information-response experiments, e.g. Cipriani & Guarino (2009), Toussaert (2017), Agranov, Dasgupta & Schotter (2020), Charness, Oprea & Yuksel (2021). See Danz, Vesterlund & Wilson (2022) for a recent evaluation of belief elicitation practices and the Online Procedures Appendix for further details on how our design introduces the elicitation method.

[17] Each part is introduced as a surprise, meaning that subjects were not informed in advance of what later parts would entail.



Round 5

If the test is POSITIVE, what is the chance that the project is a Success vs. Failure?

80 % chance the project is a SUCCESS

20% chance the project is a FAILURE

If the test is NEGATIVE, what is the chance that the project is a Success vs. Failure?

20 % chance the project is a SUCCESS

80% chance the project is a FAILURE

The test this round is Negative

The project this round is a Failure

Round	Test	Project
1	Positive	Failure
2	Negative	Failure
3	Positive	Failure
4	Positive	Success

Figure 1: Interface Screenshot at Round Five of Core Treatments

completely neglects the prior and chooses $B_{Pos} = 80$ and $B_{Neg} = 20$. Once the subject makes this selection, the outcome in this round appears at the bottom of the screen. In the example in the figure, the test was negative and the project turned out to be a failure in this round. On the right hand side of the screen, the subject can observe the signal-state realizations from all previous rounds.

The second phase encompasses rounds 101 through 200. The only difference with respect to the first phase is that subjects are asked to report their beliefs only every 10 rounds, as opposed to in every round, while receiving feedback in real time in every round. This is done to be able to assess how an additional 100 rounds of feedback would affect beliefs while keeping the experiment to a reasonable time limit.

Experimental procedures

Subjects participated in only one treatment condition (between-subjects design). Before subjects began round one, we introduced them to the belief elicitation task and the incentive-compatible BDM mechanism using simple examples. The two core treatments were conducted at the University of California, Santa Barbara and subjects (undergraduates at the university) were recruited using ORSEE (Greiner 2015). In total, 128 subjects participated (64 in each treatment).[18] The experiment, which lasted 90 minutes, was conducted using zTree (Fischbacher 2007). In addition

[18] See Online Appendix B for details on other treatments (including number of subject, location of data collection).

to the \$10 show up payment, earnings from the experiment were either \$25 or \$0, for a grand total of either \$10 or \$35.[19] Payments on average from the core treatments equaled \$22.5.

3 Results on ***Primitives*** vs. ***NoPrimitives***

We begin by confirming that initial (i.e., round one) responses in *Primitives* replicate previous findings in the literature related to BRN. We then focus on the evolution of beliefs with 200 rounds of feedback, and document differences between *Primitives* and *NoPrimitives*, first at the aggregate level and then at the individual level. These results establish that information on the primitives hinders learning from feedback such that by round 200, beliefs in *NoPrimitives* are closer to the Bayesian benchmark than beliefs in *Primitives.* We postpone analyses on the mechanisms underlying these treatment differences to the next section.

3.1 Base-rate neglect in round one of ***Primitives***

In round one of *Primitives*, the mode and the median belief reported conditional on a positive signal (B_{Pos}) is 80 percent (the pBRN prediction), which is consistent with the results for the same parameterization in Kahneman & Tversky (1972).[20] In fact, 56.3 percent of subjects in this treatment submit beliefs that are consistent with pBRN. Only 4.7 percent of subjects submit Bayesian beliefs the first time they are faced with the updating task. This share does not change if we allow for reasonable computation errors by the subjects.[21] Besides the pBRN and Bayesian benchmarks, another natural response involves signal-neglect, where beliefs conditional on either signal coincide with the prior. We find that 7.8 percent of our subjects respond in this way.

These findings confirm that the baseline condition needed for our study holds: For most subjects in *Primitives*, beliefs submitted in the first round are far from the Bayesian Benchmark. The most popular response is pBRN. We interpret this as information on the primitives inducing biased behavior (pBRN being the most prominent one).

[19] For final payment in the experiment one part is randomly selected and if the part consists of more than one decision, one decision is selected for payment in the randomly selected part. The BDM mechanism used for belief-elicitation incentives results in a binary payment of either \$0 or \$25. See Online Appendix B for details.

[20] Kahneman & Tversky (1972) only ask about beliefs conditional on a positive signal.

[21] No additional subjects are added if we let $B_{Pos} \in [36, 47]$ and $B_{Neg} \in [0, 9]$ (in percentage points).

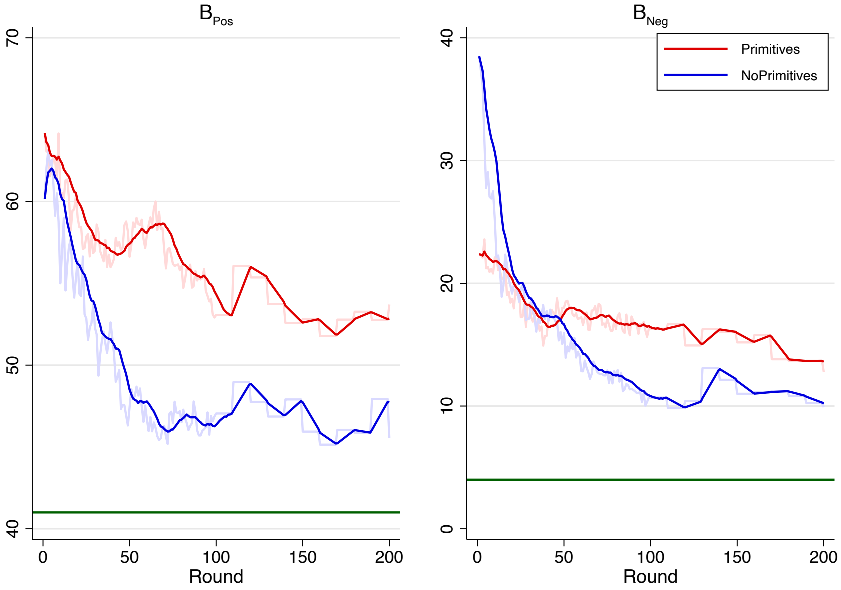


Figure 2: Evolution of Beliefs in *Primitives* and *NoPrimitives*

Notes: Shaded lines in the background show average belief in each round of each treatment. Darker lines in the foreground show beliefs averaged over prior ten rounds to make general patterns in evolution of beliefs more discernible. The horizontal green lines correspond to the Bayesian benchmark.

3.2 Learning in *Primitives* vs. *NoPrimitives*

Figure 2 presents the evolution of beliefs, B_{Pos} andB_{Neg}, at the aggregate-level across all rounds in *Primitives* vs. *NoPrimitives*.[22] While beliefs for both treatments start far from the Bayesian benchmark and move towards this benchmark, after 200 rounds beliefs in *NoPrimitives* are closer to it, and most of the adjustment occurs in the first 100 rounds.

Specifically, average beliefs in *Primitives* move from $(B_{Pos}, B_{Neg}) = (64, 22)$ in round one to $(53, 16)$ in round 100. At this point, average beliefs are still twelve percentage points away from the Bayesian benchmark conditional on either signal. Note, however that there could be many factors that slow down learning in *Primitives*. The *NoPrimitives* treatment serves as a natural benchmark allowing us to contextualize results from *Primitives*. In *NoPrimitives*, average beliefs in round one are equal to $(60, 39)$, which is quite far from the Bayesian benchmark. Yet after 100 rounds beliefs move close to the benchmark, reaching $(47, 11)$.

To provide statistical analysis on the differences between *Primitives* and *NoPrimitives* depicted in Figure 2, we focus on two questions: (1) Are there treatment differences in how far beliefs are

[22] On average, subjects will experience 29 (59) rounds with a positive and 71 (141) rounds with a negative signal by the end of 100 rounds (200 rounds).

to the Bayesian benchmark? (2) Are beliefs different between the two treatments?[23,24]

For question (1), we use distance to Bayesian benchmark: $|B_j - B_j^{Bay}|$ for $j \in \{Pos, Neg\}$, corresponding to the absolute value of the deviation from the benchmark. For question (2), we directly use B_{Pos} and B_{Neg}. To determine statistical significance, we run regressions where the left hand side variable is the measure relevant to the question and the right-hand side variable is a treatment dummy. [25] Such analysis reveals beliefs in *NoPrimitives* to be significantly closer to the Bayesian benchmark relative to beliefs in *Primitives* by round 100 (p-value 0.011), a finding that does not change after 200 rounds (p-value 0.007). Furthermore beliefs are different between the two treatments (p-value 0.056 in round 100, p-value 0.049 in round 200).

3.3 Heterogeneity

To provide an overview of the heterogeneity in responses, Figures 3 and 4 present the distribution of beliefs in *Primitives* and *NoPrimitives* at the initial and final rounds. As mentioned earlier, most subjects (56.3 percent) submit beliefs consistent with pBRN in round one of *Primitives.* By round 200, however, the distribution of beliefs in *Primitives* has shifted significantly, with one large cluster close to or at the pBRN point and another one close to or at the Bayesian point. In fact, 13 percent of subjects submit beliefs consistent with pBRN in both round one *and* round 200.[26]

For *NoPrimitives*, subjects' beliefs in round one can largely be organized into two groups. A large mass of subjects (forty-five percent) submit $(B_{Neg}, B_{Pos}) = (50, 50)$. This is consistent with subjects recognizing that they have no information to base these beliefs on (since they have not been given the primitives). Another large group of subjects (forty-eight percent) submit beliefs that suggest they consider the labels we used for the signals (positive vs. negative) to provide some information, i.e., $B_{Pos} > B_{Neg}$. By round 200 (right plot of Figure 4), the mass at $(50, 50)$ largely disappears and fifty-nine percent of subjects are at ±10 percentage points of the realized frequencies.

These patterns suggest long-run differences between *Primitives* and *NoPrimitives* to be possibly

[23] Note that (1) are (2) are related, but conceptually different questions. For example, beliefs can be different in the two treatments while being equally distant from the Bayesian benchmark (resulting from deviations in opposite direction).

[24] In Online Appendix C.1, following an approach first introduced by Grether (1980), we also report treatment differences in aggregate measures of base-rate neglect by focusing on changes in log likelihood ratios.

[25] We estimate a system of equations using seemingly unrelated regressions. The p-values that we report to evaluate treatment effects result from using a Wald test on the hypothesis that both treatment coefficient estimates (focusing on B_{Pos} and B_{Neg}) are equal to zero. See Online Appendix C.1 for further details.

[26] By round 200, 34 percent of subjects are at ±10 percentage points of the pBRN benchmark and a similar proportion (36 percent) is within ±10 percentage points of the realized frequencies.

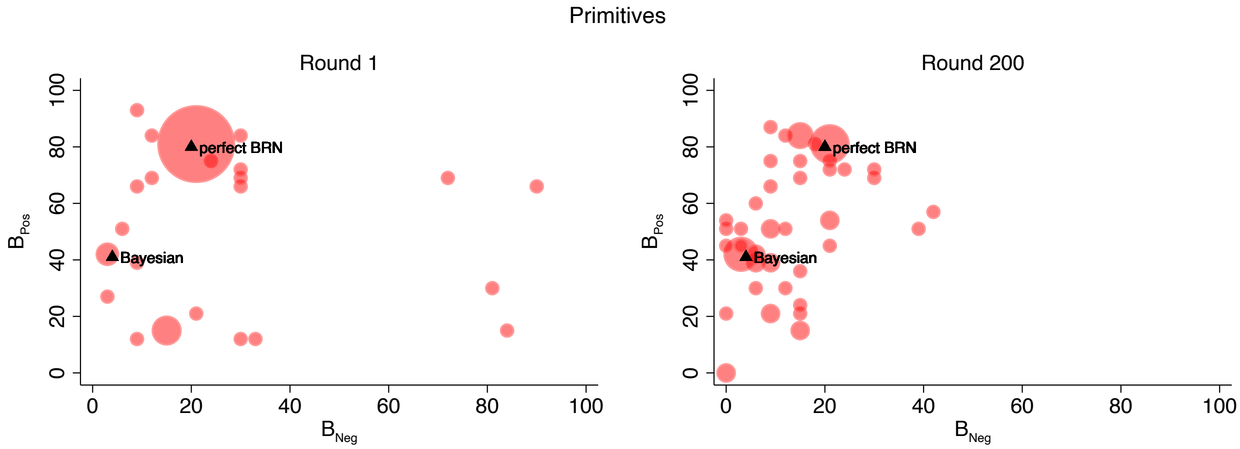


Figure 3: Density plots for *Primitives*

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Beliefs are presented at the individual level (rounded to multiples of 3). The size of each bubble represents the number of subjects with such beliefs.

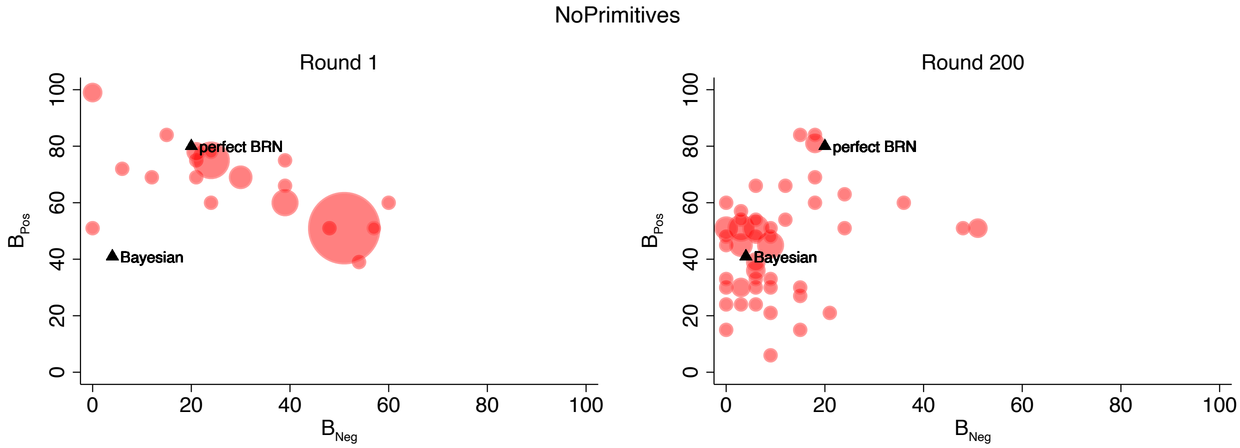


Figure 4: Density plots for *NoPrimitives*

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Beliefs are presented at the individual level (rounded to multiples of 3). The size of each bubble represents the number of subjects with such beliefs.

driven by those subjects who initially display perfect base-rate neglect in *Primitives*. To begin to assess this possibility, we divide subjects in *Primitives* into two types: those who submit the pBRN beliefs in round one and all others. In Figure 5 we depict the average evolution of beliefs for these two types and compare them to the beliefs of subjects in *NoPrimitives*. The long-run beliefs of round one pBRN subjects are different from subjects in *NoPrimitives*. For example, there is a fifteen percentage-point difference in the average B_{Pos} between the two groups by round 200. Long-run beliefs of these subjects are significantly different (p -value 0.001) and farther away from the

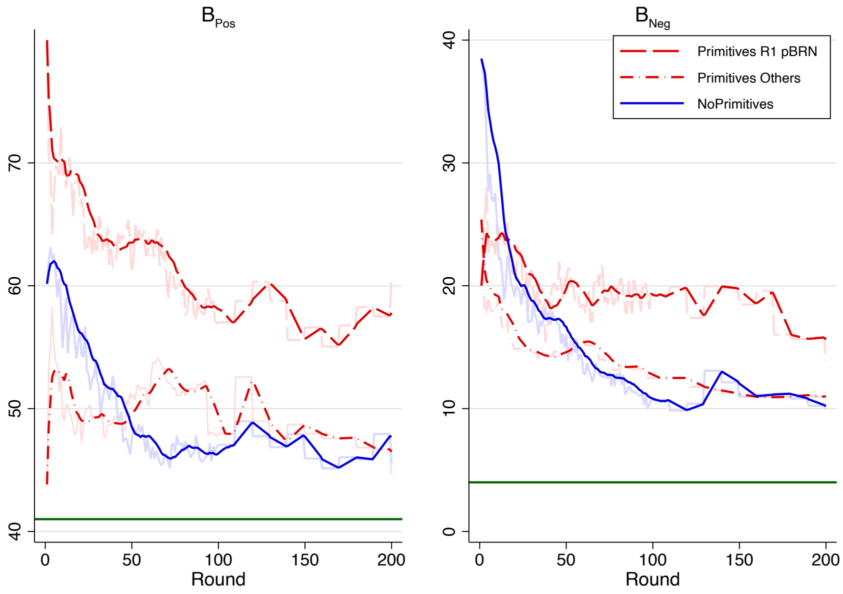


Figure 5: Evolution of Beliefs for R1 pBRN Subjects and Others in *Primitives* vs. *NoPrimitives*

Notes: Shaded lines in the background show average belief in each round of each treatment. Darker lines in the foreground show beliefs averaged over prior ten rounds to make general patterns in evolution of beliefs more discernible. The horizontal green lines correspond to the Bayesian benchmark. Beliefs are separated by round one behavior. *Primitives R1 pBRN* denotes beliefs of subjects who start at the pBRN point. *Primitives Others* refers to others in the same treatment.

Bayesian benchmark (p -value < 0.001) relative to subjects in *NoPrimitives*. The average belief of all others (i.e., non-pBRN subjects), however, are not different from the average beliefs of subjects in *NoPrimitives* (p -value 0.760).[27]

We can further split (round one) pBRN subjects into those who are still stuck at the pBRN response in round 200 and those who are not. By round 200, beliefs of those who are not stuck at the pBRN response are still significantly farther from the Bayesian benchmark compared to subjects in *NoPrimitives* (p -value 0.022).[28] This suggests that both kinds of pBRN subjects (including those who revise their beliefs away from the pBRN response) are responsible for hindering learning.[29] Overall, these mechanical effects suggest that the *Primitives* treatment operates by inducing certain initial misconceptions, and that, of all misconceptions, it is principally those that induce pBRN beliefs in round one that hinder learning from feedback.

[27] Despite similarity in long-run beliefs between these groups, we do not want to suggest that others in *Primitives* behave exactly the same as those in *NoPrimitives*. As seen in Figure 5, others in *Primitives* learn faster, suggesting that they are using both data and information on primitives to learn.

[28] Tables 12 to 14 in Online Appendix G provide additional details of this comparison.

[29] We see evidence of both smooth changes in beliefs (consistent with Bayesian updating with an initially incorrect answer) and of sudden large shifts that occurs only after sufficient evidence accumulates (consistent with models of hypothesis testing, as in Ortoleva (2012)). However, our experiment was not designed to distinguish between different learning models, but rather to focus on long-run outcomes and persistence of mistakes.

Result #1: *Long-run beliefs in NoPrimitives are different, and closer to the Bayesian benchmark, than beliefs in Primitives. This treatment effect vanishes when we exclude subjects with pBRN beliefs in round one of Primitives.*

4 Mechanisms

In this section, we investigate possible mechanisms underlying the treatment differences between *Primitives* and *NoPrimitives*. First, it is possible that subjects in *Primitives*, particularly those who are giving the pBRN response, have formed an understanding of the environment (based on information on the primitives) that incorrectly justifies and makes them more confident in their initial response. Here, we use the term "confidence" to capture how strong the agent's prior beliefs are about the optimality of their responses in round one. The degree to which subjects' beliefs will change with new information (available through feedback) will depend on the strength of their prior. Thus, a reasonable first hypothesis on why subjects don't learn as much in *Primitives* is that the additional information provided to them in this treatment makes them more confident in their (incorrect) initial responses, and hence less responsive to new information.

A second mechanism, closely tied to the first, builds on the hypothesis that subjects in *Primitives* could be highly confident in their initial responses. Confidence in one's initial response can impact how attentive subjects are to the feedback. A strong prior decreases incentives to engage in costly learning. It is possible that subjects in *Primitives* don't learn as much because they choose to engage less with the feedback relative subjects in *NoPrimitives*.

The impact of these two mechanisms crucially depends on learning being costly. Note that while we designed the experiment to make learning from feedback quite easy (by making it available at any point), subjects still must pay some cost to process the many rounds of feedback they receive to be able to learn from it. This suggests that lowering the cost of learning can improve optimality of long-run beliefs.

In this section, we report results on additional treatments that allow us to assess the importance of initial confidence, attention, and costly learning.

4.1 Confidence

If confidence in an incorrect initial answer is the reason why subjects don't learn as effectively in *Primitives*, then a shock to their confidence should facilitate learning. To test this possibility, we conduct a new treatment, *Primitives w/ shock*, that is identical to *Primitives* except for one difference: If a subject submits an incorrect answer in round one, the computer interface sends them a message that says that their answer is incorrect before they start with round two.[30]

Given round one responses, 90 percent of subjects in *Primitives w/ shock* received a message that stated both of their initial answers (on B_{Neg} or B_{Pos}) were incorrect.[31] Figure 6 depicts the evolution of beliefs in *Primitives w/ shock* using an orange line. The figure also includes *Primitives* and *NoPrimitives* (red and blue lines, respectively) for comparison. The figure reveals that long-run beliefs (round 200) are different between *Primitives w/ shock* and *Primitives* (p-value 0.013), and closer to the Bayesian benchmark in *Primitives w/ shock* relative to *Primitives* (p-value 0.021). The differences are most striking for beliefs conditional on a positive signal. For example, there is a sharp contrast between *Primitives w/ shock* and *Primitives* in how much B_{Pos} changes in the first 50 rounds. Overall, the gap between the two treatments (between the orange and the red line) widens with experience. By contrast, particularly after the first 50 rounds, beliefs in *Primitives w/ shock* are very similar to beliefs in *NoPrimitives.* Table 10 in Online Appendix D provides further statistical analysis supporting these observations.

Result #2: *Shocking confidence of subjects in their initial response (by telling them their answers are incorrect) improves optimality of beliefs. Long-run beliefs in Primitives w/ shock are not different from those in NoPrimitives.*

It is also important to note that, in contrast to our findings in *Primitives*, subjects who display perfect BRN in round one of *Primitives w/shock* learn as well as others in the same treatment. Average beliefs in round 200 for these subjects (who display perfect BRN in round one) are 45 for B_{Pos} and 12 for B_{Neg}. The corresponding values are 41 and 11 for others in the same treatment. These differences are not statistically significant (p-value 0.598).[32] These patterns in *Primitives w/ shock* confirm that all subjects, including those who start at the pBRN point, are capable and willing to learn from feedback when they are informed about the incorrectness of their initial response. These results rule out the possibility that pBRN subjects are intrinsically worse at learning from feedback compared to others, and further supports the hypothesis that initial confidence in the pBRN response is driving the treatment differences between *Primitives* and *NoPrimitives.*

[30] Specifically, subjects were told either both of their answers (on B_{Pos} or B_{Neg}) were incorrect, or at least one of their answers were incorrect. In particular, subjects who submitted a Bayesian response to both questions didn't receive any message.

[31] In addition, three percent of subjects received a message indicating that at least one of their answers were incorrect. In Online Appendix D, we document that the results in one round of *Primitives w/ shock* are not statistically different from those of *Primitives*, which is to be expected since the treatments are identical up to the end of round one.

[32] Figure 25 in Online Appendix G reproduces Figure 5 depicting the evolution of beliefs in *Primitives w/shock* separately for (round one) pBRN subjects vs. others. This appendix also includes further analysis on differences with respect to these types.

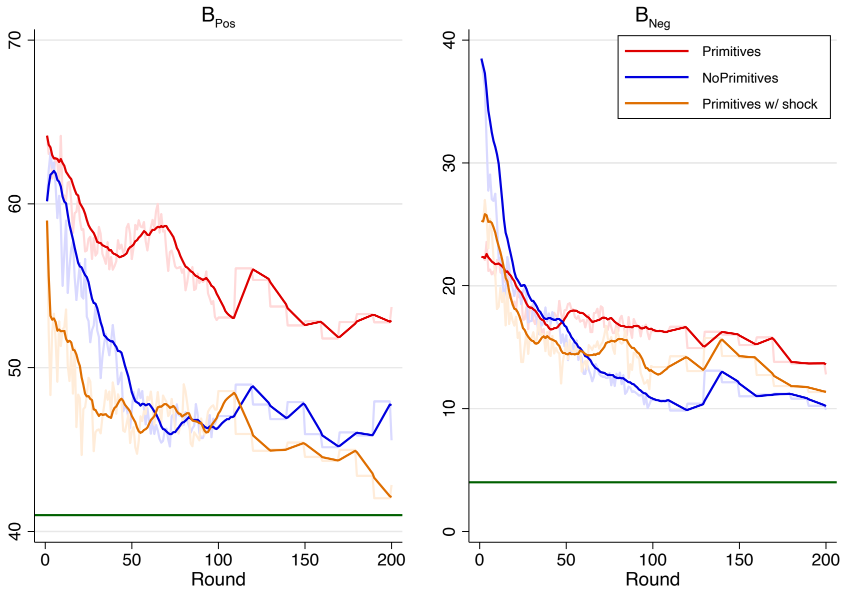


Figure 6: Comparing Evolution of Beliefs in *Primitives w/ shock* to *Primitives* and *NoPrimitives*

Notes: Shaded lines in the background show average belief in each round of each treatment. Darker lines in the foreground show beliefs averaged over prior ten rounds to make general patterns in evolution of beliefs more discernible. The horizontal green lines correspond to the Bayesian benchmark.

4.2 Attentiveness

There are two ways in which confidence in initial (round one) responses may hinder learning from feedback. First, confidence can lead a subject to put more weight on her initial answer relative to new information or feedback. Second, confidence can lead a subject to pay less attention and engage less with feedback. In this section, we introduce new treatments to assess differences in attentiveness between *Primitives* and *NoPrimitives*.

In the original experiment, feedback was visually available to the subjects at any point at almost no cost. But, given the stochastic nature of the task, no single round of feedback can invalidate a subject's beliefs. With attentiveness, we mean to capture a more meaningful notion in which subjects don't just look at the data but also engage with it in a way that could effectively change their beliefs. For example, the empirical distribution of the state conditional on each signal after 100 rounds provides a strong statistical signal that the pBRN response is not correct. While the data underlying this signal is readily available, subjects might not sufficiently engage with the data in this way, potentially because confidence in their initial answers endows little value to such an exercise. This is precisely the type of inattentiveness we hope to capture in the new experiment.

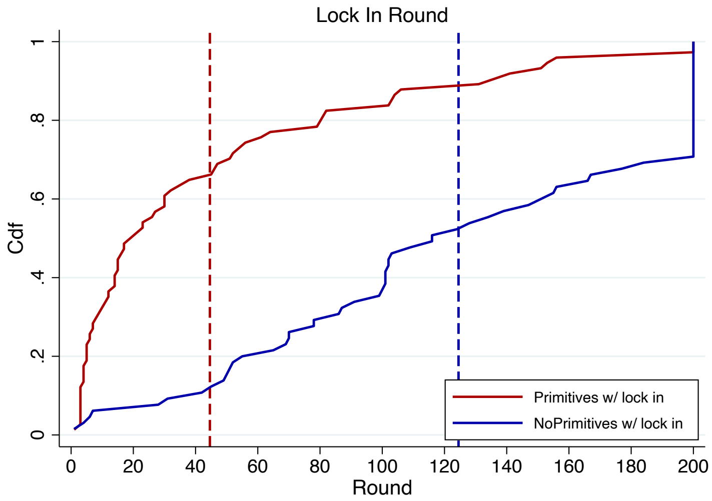


Figure 7: Distribution of Lock In Decisions

Notes: Subjects who never locked in are coded as locked in at round 200. Vertical lines denote mean values. Vertical dashed lines indicate mean value by treatment.

Studying the degree to which learning is slowed down by partial attentiveness to the feedback is difficult because it is not possible to directly observe attentiveness (as defined above) in our core treatments. To overcome this challenge, we run two diagnostic treatments, *Primitives w/ lock in* and *NoPrimitives w/ lock in*. These treatments are identical, respectively, to the main parts of *Primitives* and *NoPrimitives* (as described and analyzed earlier) except for one difference in how subjects move through the 200 rounds of feedback. Critically, subjects are allowed, in the new treatments, to "lock in" their choices at any round, which automatically implements their latest responses for all future remaining rounds.[33] We do not take the lock-in round as a perfect measure of attentiveness, but we interpret differences between the *Primitives w/ lock in* and *NoPrimitives w/ lock in* in terms of lock in decisions to reflect differences between these two environments in willingness to engage with the feedback.

Figure 7 shows the cumulative distribution of round of lock in decisions in *Primitives w/ lock in* and *NoPrimitives w/ lock in*.[34] There are large differences between these two treatments with respect to willingness to engage with the feedback. In fact, the distribution of lock-in decisions in *NoPrimitives w/ lock in* first-order stochastically dominates that of *Primitives w/ lock in*.[35] In

[33] Instructions indicated clearly that subjects wouldn't be able to leave the experiment earlier by locking-in their responses. Thus, we removed incentives to use the lock in option to end the experiment earlier.

[34] In Online Appendix E, we confirm that initial responses are similar between the core treatments and the new ones with the lock in option. One difference is that there are slightly fewer pBRN subjects in *Primitives w/ lock in* relative to *Primitives*: 42 percent vs. 56 percent (p -value 0.093). As is clear from the stark treatment differences in lock in choices, this does not impact the conclusions that we can draw from the lock-in treatments.

[35] We test for first-order stochastic dominance using the test in Barrett & Donald (2003). The test consists of two

Primitives w/ lock in, only half the subjects choose to see more than 20 rounds of feedback and only four percent of subjects choose to see all rounds of feedback. By contrast, in *NoPrimitives w/ lock in*, 94 percent of subjects choose to see more than 20 rounds of feedback and 31 percent of subjects choose to see all rounds of feedback. The average lock-in round is roughly three times higher in *NoPrimitives w/ lock in* (difference p -value < 0.001).

Result #3: *Subjects lock in their choices earlier in Primitives w/ lock in relative to NoPrimitives w/ lock in.*

Interestingly, the average lock-in round is not very different between (round one) pBRN subjects and others in *Primitives w/ lock in*, with both types engaging less with data relative to subjects in *NoPrimitives w/ lock in* (p-value < 0.001 for both types).[36] But the reasons why subjects don't engage as much with data are likely to be different for pBRN subjects and others. For some pBRN subjects, confidence in their initial model may make them reluctant to engage with data. For others or those who are more willing to question their model, having access to the primitives means they can learn more effectively relative to subjects in *NoPrimitives*, thus requiring less rounds of feedback. In fact, when we compare long-run beliefs, we find once again that learning is hindered for (round one) pBRN subjects in *Primitives w/ lock in* while there are essentially no differences in learning between others in *Primitives w/ lock in* and subjects in *NoPrimitives w/ lock in.*[37]

Overall, these treatments suggest important differences between the two environments corresponding to our core treatments (with and without primitives) in willingness to engage with and learn from feedback. Hence, these results are in support of our hypothesis that differences in attentiveness to feedback are an important factor in explaining differences in long-run beliefs between *Primitives* and *NoPrimitives*.

4.3 Costly attention

Learning from feedback requires engaging with that feedback in a way that may be costly. In this section, we investigate the extent to which learning costs play a role in hindering learning. We run

steps. We first test the null hypothesis that the distribution in *NoPrimitives w/ lock in* either first order stochastically dominates or is equal to the distribution in *Primitives w/ lock in.* We reject this null hypothesis (p-value <0.001). We then test the null hypothesis that the distribution in *Primitives w/ lock in* first order stochastically dominates the distribution in *NoPrimitives w/ lock in.* We cannot reject the null in this case (p-value 0.827).

[36] Specifically, in *Primitives w/ lock in*, (round one) pBRN subjects lock in slightly later than others (p-value 0.079). The difference is only marginally significant if we take out (round one) Bayesian subjects from others (p -value 0.097). It is worth noting that there are 12 subjects (39 % of pBRN subjects) in this treatment who remain at the pBRN response for *all* 200 rounds, but their average lock-in round is 61.

[37] Figure 19 in Online Appendix E reproduces Figure 2 for these new treatments. Figure 26 in Online Appendix G reproduces Figure 5 separating behavior for (round one) pBRN subjects and others.

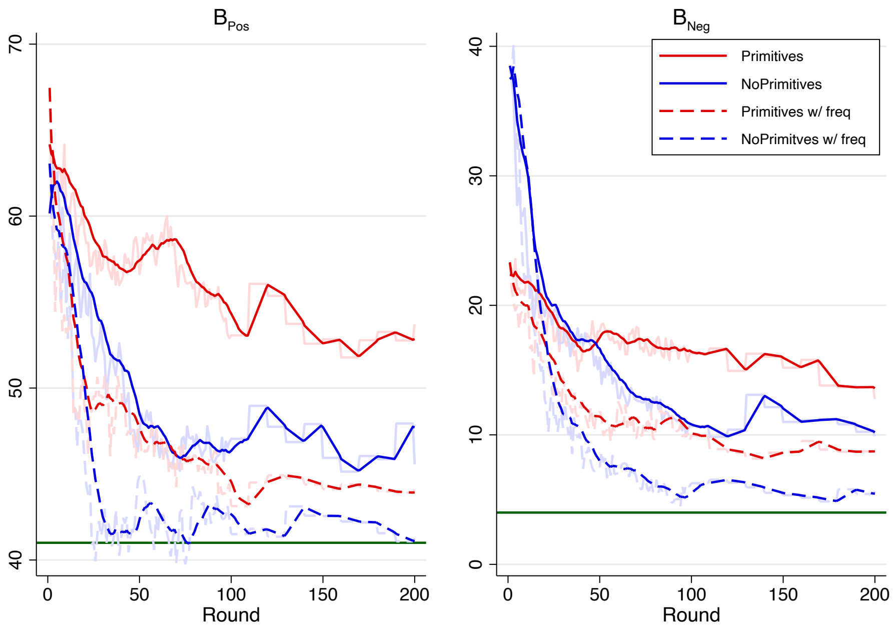


Figure 8: Evolution of Beliefs in Treatments with Frequencies Relative to Core Treatments

Notes: Shaded lines in the background show average belief in each round of each treatment. Darker lines in the foreground show beliefs averaged over prior ten rounds to make general patterns in evolution of beliefs more discernible. The horizontal green lines correspond to the Bayesian benchmark.

two new treatments, labeled as *Primitives w/ freq* and *NoPrimitives w/ freq*. These treatments are identical, respectively, to the main parts of *Primitives* and *NoPrimitives* (as described and analyzed above) except for one difference in how the feedback is presented to the subjects. Recall that, in the earlier treatments, subjects were provided feedback on a round-by-round basis and feedback from all previous rounds were recorded in a history table (see Figure 1). In *Primitives w/ freq* and *NoPrimitives w/ freq*, we still provide feedback on a round-by-round basis. But feedback from all previous rounds is now aggregated and presented in a two-by-two table which summarizes the total number of actual rounds in which each combination of the signal and state realization were observed. In addition, we also compute empirical frequencies. For example, we report to subjects the total number of rounds in which they observed the signal to be positive in the past and the empirical frequency of success among these rounds.[38] The goal of these new treatments is to minimize the cost of attentiveness to feedback.

Figure 8 depicts the evolution of beliefs with feedback in the treatments with frequency in-

[38] Figure 20 in Online Appendix F provides a screenshot from this treatment. To ensure that subjects are indeed aware of all this information presented to them, the interface also requires subjects to give us back the frequency information (which is presented on the same screen) every 20 rounds. For details see Online Appendix B.

formation and contrasts these to the core treatments.[39] The figure reveals stark differences in learning when feedback is presented in an aggregated form. By round 200, beliefs in the treatments with frequency information are different from those in the core treatments (p-value is 0.010 between *Primitives w/ freq* and *Primitives*, and 0.033 between *NoPrimitives w/ freq* and *NoPrimitives*) and closer to the Bayesian benchmark relative to core treatments (p-value < 0.001 for both comparisons). The evidence also suggests convergence in behavior between the treatments with frequency information. While Figure 8 reveals different learning dynamics in these treatments (with the dashed red line depicting *Primitives w/ freq* consistently hovering above the dashed blue line depicting *NoPrimitives w/ freq*), long-run differences observed in our core treatments (between *Primitives* and *NoPrimitives*) are greatly reduced in new treatments (between *Primitives w/ freq* and *NoPrimitives w/ freq*). By round 200, beliefs are not statistically different between *Primitives w/ freq* and *NoPrimitives w/ freq* (p-value 0.196), and not statistically different with respect to distance to Bayesian benchmark (p-value 0.313). [40]

To summarize, we find that eliminating costs associated with attending to the data, by presenting feedback in terms of empirical frequencies, significantly improves optimality of long-run behavior. This is true regardless of whether subjects were provided information on the primitives or not. This suggests attention costs play an important role in hindering learning in both *Primitives* and *NoPrimitives.*

4.4 A model of learning

We have established that confidence in an initially incorrect answer can negatively impact the optimality of long-run behavior for two related reasons: Subjects place more weight on a stronger prior, and subjects are less attentive to feedback that is costly to process. At this point we would like to assess the relative importance of prior strength and attentiveness, since these mechanisms have different policy implications regarding how to correct biases.

Consider the following counterfactual: Suppose that subjects in *Primitives*, with their presumably stronger priors, were equally attentive to feedback as subjects in *NoPrimitives.* By how much

[39] In Online Appendix F we provide a more detailed analysis of treatment comparisons. Table 11 of this appendix summarizes statistical analysis presented in this section. In particular, we show that the new treatments, *Primitives w/ freq* and *NoPrimitives w/ freq*, do not differ, respectively, from *Primitives* and *NoPrimitives* in terms of round one behavior.

[40] There is some evidence to suggest that the difference in long-run beliefs between *Primitives w/ freq* and *NoPrimitives w/ freq* are driven by those subjects in the former treatment who are consistent with pBRN in round one. Despite the frequency information, eight percent of subjects in this treatment are consistent with pBRN both in rounds one and 200. See Online Appendix G for more analysis, including a reproduction of Figure 5 for these treatments.

would the gap in distance to the Bayesian benchmark between the two treatments be reduced? Because attention is not directly observable in our core treatments, to answer this question we will rely on a simple learning model.

We assume the agent is uncertain about the true likelihood p of an event (e.g., the project being a success conditional on a positive signal). The agent's prior is given by the Beta distribution and is characterized by two parameters p_0 and η, such that:

$$\mathbb{E}(p \mid p_0, \eta) = p_0 \quad \text{and} \quad \mathbb{V}(p \mid p_0, \eta) = \frac{p_0(1-p_0)}{\eta+1}.$$

While p_0 denotes the expected value of p, η captures the strength of the prior and, hence, can be interpreted as a measure of the agent's confidence.[41]

The agent updates beliefs on p using outcomes from a Bernoulli process where the probability of the event happening is the true p. The data observed by the agent can be characterized by two parameters: the number of observations n, and the observed frequency of the event among these observations f. Partial attentiveness can be introduced naturally here by assuming that the agent remembers only a subset of the observations. To keep things simple, we model this by assuming the agent misremembers n as σn for some $\sigma \in [0,1]$ (but remembers f correctly).[42] The agent's updated posterior is still characterized by a Beta distribution with adjusted parameters $\tilde{p}$ and $\tilde{\eta}$:

$$\tilde{p} = \left(\frac{\eta}{\tilde{\eta}}\right) p_0 + \left(1 - \frac{\eta}{\tilde{\eta}}\right) f \quad \text{and} \quad \tilde{\eta} = \eta + \sigma n \tag{1}$$

In summary, the model describes how beliefs evolve with feedback as a function of three parameters: p_0, prior expected value on p; η, a measure of initial confidence; and σ, attentiveness to data.

We assume that the agent's reported belief corresponds to the expected value of p as described above. In our data, we directly observe the feedback experienced by subjects (n and f). Prior expected value (p_0) can be directly identified from initial responses. However, since the evolution of beliefs depend on σ/η, we need a way to separately identify these parameters.[43] We do so by using the treatments with frequency information. Specifically, we estimate η from the the treatments with frequency information by assuming that attentiveness to data is maximal, i.e., $\sigma = 1$. Then, taking as given the estimated values of η (from the new treatments), we use data from the core treatments

[41] In the standard formulation, the Beta distribution is characterized by two parameters: α, β such that $\mathbb{E}(p \mid \alpha, \beta) = \frac{\alpha}{\alpha+\beta}$ and $\mathbb{V}(p \mid \alpha, \beta) = \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$. The mapping to p_0 and η are such that $p_0 = \frac{\alpha}{\alpha+\beta}$ and $\eta = \alpha + \beta$.

[42] The model could be enriched by assuming that the agent remembers each observation independently with probability σ. In expectation, the agent will misremember n as σn and f as f. Since our estimation will focus on aggregate results, we simplify the model by eliminating the randomness around this.

[43] By Equation 1, expected beliefs change with observed frequency f as a function of $\frac{\eta}{\tilde{\eta}} = \frac{\eta}{\eta+\sigma n} = \frac{1}{1+\frac{\sigma}{\eta}n}$.

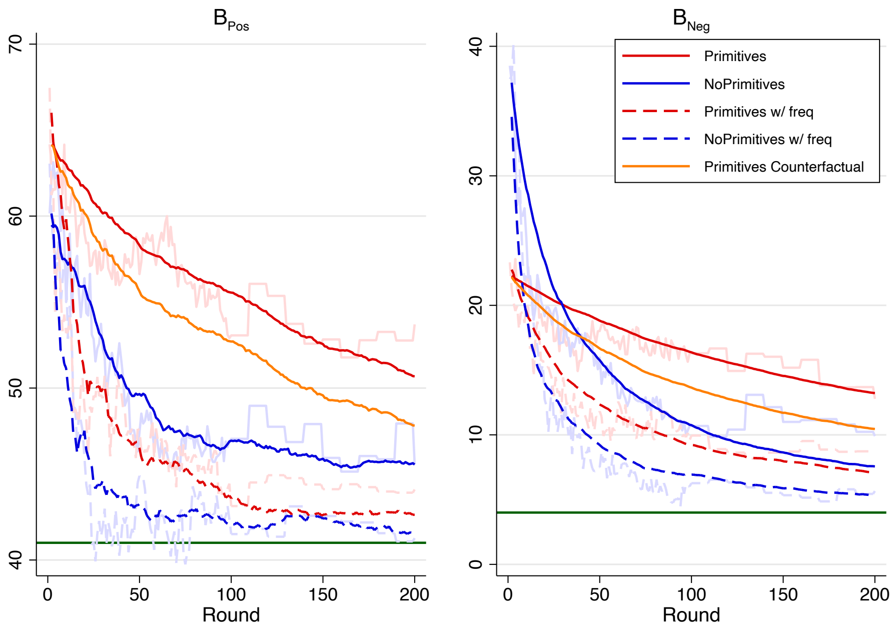


Figure 9: Estimates of the Learning Model for Treatments with Frequencies and Core Treatments

Notes: Shaded lines in the background show average belief in each round of each treatment. Darker lines depict estimates from the learning model. Orange line represent a counterfactual estimate where subjects in *Primitives* are set to be as attentive as those in *NoPrimitives* (keeping confidence level the same). The horizontal green lines correspond to the Bayesian benchmark.

to estimate σ.[44]

Figure 9 plots the model predictions overlaid on actual data. We find that the model (using only a few parameters) does a remarkable job capturing the qualitative differences between the treatments in terms of how beliefs change with feedback. Focusing on the treatments with frequency information (depicted using dashed lines), differences in speed of learning are attributed to differences in confidence. Specifically, our estimates for η are substantially higher for those subjects who were given the primitives vs. those who were not.[45]

Nonetheless, our estimates for σ reveal that there are also important differences between *Primitives* and *NoPrimitives* in terms of attentiveness to feedback. While subjects in both treatments are extracting less information from the feedback than those in the treatments with frequency in-

[44] We use least squares estimation to fit average behavior in each treatment. In Online Appendix F, we present the details of the estimation procedure as well as results from an alternative estimation where we also account for heterogeneity across subjects. This analysis generates the same qualitative conclusions about the importance of the two channels discussed above.

[45] Estimates of η for B_{Pos} (B_{Neg}) are 4.2 and 2.2 (5.9 and 25) in *Primitives* and *NoPrimitives*, respectively. Statistical tests using bootstrapping show differences to be significant (p-value < 0.001 for both B_{Pos} and B_{Neg}).

formation, our estimates for σ are higher (for both B_{Pos} and B_{Neg}) in *NoPrimitives* relative to *Primitives*.[46]

These results indicate that both channels—confidence and attentiveness to feedback—play an important role in determining how much subjects learn from their experiences. But, it remains an open question, how much subjects in *Primitives* could have learned (keeping confidence in their initial response constant) if they had been as attentive as those in *NoPrimitives*. The learning model allows us to compute this counterfactual, which is included (with an orange line) in Figure 9. This exercise leads to the following observations. For low levels of feedback (early rounds), differences between *Primitives* and *NoPrimitives* are primarily driven by differences in confidence and differences in starting points. This is revealed by the proximity of the orange line to the red line in this region. But, as the amount of feedback increases (as we move towards 200 rounds), the orange line departs substantially from the red line. This suggests that, in the long run, differences in attentiveness between the two treatments also play a significant role in explaining the differences in beliefs.

Result #4: *Beliefs move closer to the Bayesian benchmark when feedback is presented in a processed way. Results from a learning model suggest differences in long-run beliefs between our core treatments (Primitives vs. NoPrimitives) to be driven by differences in both confidence and attentiveness.*

4.5 Transfer learning: Behavior with different primitives

So far our design does not distinguish between two different ways in which subjects who know the primitives but initially submit incorrect beliefs, can learn from feedback. The first involves subjects simply adjusting their beliefs to be consistent with the data. The second entails a deeper form of learning, where subjects gain an understanding of why their initial answers were incorrect (for example, that they failed to account for the base-rate).[47]

We tackle the question of what subjects are learning from their experiences in the last part of our core treatments. In this part, subjects face a new updating task in which the primitives are changed to $p' = .95$ and $q' = .85$. Prior to this part, we presented subjects in these treatments with ample feedback processed for them such that almost all subjects converged to beliefs very close to

[46] Estimates of σ for B_{Pos} (B_{Neg}) are 0.10 and 0.18 (0.19 and 0.35) in *Primitives* and *NoPrimitives*, respectively. Statistical tests using bootstrapping show differences to be significant (p-value < 0.001 in both cases).

[47] A few papers have studied transfer of learning across environments and find limited evidence for it (e.g. Kagel (1995), Cooper & Kagel (2009), Cooper & Van Huyck (2018)). In Esponda et al. (2023*a*), we provide a more detailed discussion of the literature on transfer learning and examine the related, but different, question of whether subjects can learn not to update in the wrong direction when they know the primitives of the problem.

the Bayesian benchmark.[48] In this last part of the experiment, subjects in the core treatments are asked to report beliefs just once, without any feedback. Note that subjects in both the *Primitives* and *NoPrimitives* treatments are now given the primitives of this new updating task, but only subjects in *Primitives* could have learned to take the base rate into account from their experience in the original task.

Our main finding is that the treatment effect switches direction relative to earlier parts, and now subjects in *Primitives* are both closer to the Bayesian benchmark (p-value 0.022) and exhibit a much lower rate of base-rate neglect relative to *NoPrimitives.* For example, if we allow for ± 5 percentage points in each belief, then 47 percent of subjects in *NoPrimitives* and 25 percent of subjects in *Primitives* are classified as pBRN, i.e., $(B_{Pos}^{pBRN'}, B_{Neg}^{pBRN'}) = (85, 15)$. The results suggest that at least some subjects in *Primitives* can extrapolate from what they learned with the baseline primitives to new primitives. However, we should also note that such learning is partial as average beliefs in *Primitives*, $(B_{Pos}, B_{Neg}) = (85, 41)$, continue to be far from the Bayesian benchmark, $(B_{Pos}^{Bay'}, B_{Neg}^{Bay'}) = (99, 77)$.[49]

Result #5: *When subjects encounter a new updating task with new primitives, beliefs in Primitives are closer to Bayesian benchmark than those in NoPrimitives. This suggests that some subjects in Primitives learn to take the prior into account.*

5 Evidence beyond the updating problem

An important question motivating this paper is whether systematic biases in decision making are self-corrected in the long run when agents are accumulating feedback informative of optimal behavior. Our paper establishes a negative answer to this question in a specific setting where the dominant deviation from optimal behavior is base-rate neglect. In this section, we provide evidence on the generalizability of these results to other settings.

The results presented in Section 4 suggest that failures of learning in our original experiment, as captured by the long-run difference between *Primitives* and *NoPrimitives*, are driven by confidence in an incorrect initial answer. Confidence hinders learning in two ways: (i) makes subjects less responsive (put less weight) on new information, (ii) lowers attentiveness to such information. These findings provide insights on what other types of mistakes might fail to be self-corrected with

[48] See Online Appendix B for more details on the implementation and Online Appendix C.2 for the results. These results are consistent with findings in Fudenberg & Peysakhovich (2016). The paper studies an environment with adverse selection and shows that subjects tend not to use feedback optimally. However, processing the same data for subjects by presenting simple averages gets individuals most of the way to optimality.

[49] Figure 28 in Online Appendix H presents the distribution of beliefs in both treatments for this round.

experience. Our results suggest that mistakes that are driven by an incorrect understanding of the environment that misses or misrepresents some aspects of reality might not be corrected. Our use of the term *incorrect mental model* is intended to capture any misconception that produces suboptimal behavior while inducing confidence in such behavior.

Not all mistakes are driven by incorrect mental models, as we have just defined. Mistakes also arise when it is cognitively costly to identify optimal behavior. These costs could include everything from comprehension of primitives of the problem to using these primitives to make an inference about optimal action. To lower costs, an agent might use simpler (cognitively less costly but suboptimal) methods to determine which action to take. In such cases, the agent will be self-aware of the possibility of making a mistake, will be less confident in their initial answer, and open to correcting their behavior when there is new information provided that is indicative of optimal behavior.

In different words, our results suggest the following hypotheses. First, in settings in which agents have confidence on choices that are actually suboptimal, learning will be hindered. Meanwhile, in cases where subjects are aware of a possible mistake, they would have lower confidence in their initial answer and increase engagement with data.

We conduct four more treatments, in a new setting, to provide a first test of these ideas.[50] The specific problem we use is a variation of the problem studied in Ali, Mihm, Siga & Tergiman (2021). The agent and a computerized player simultaneously vote either for an option that pays \$6 for sure (option 1), or for an option that pays either \$0 or \$10 (option 2). Option 1 determines the agent's payoff if there is one or more votes for it. Option 2 is selected only if it gets both votes. Option 2 pays \$10 whenever a random integer in $\{1, ..., 100\}$ (uniformly selected) is higher than 60. The agent knows that the computer is programmed to vote for option 2 whenever the random number is higher than 60. While there is an appearance of a safe (option 1) vs. risky (option 2) choice, voting for option 2 is actually dominant. The computer's vote carries information since the computer votes for option 2 only when option 2 pays \$10. If the subject votes for option 2, her payoff will be either \$6 (when the computer votes for option 1) or \$10 (when the computer votes for option 2). However, to realize the dominance of voting for option 2, the agent has to reason contingently, focusing on the event when their vote is pivotal.[51] Subjects who fail to do so might incorrectly perceive this as a choice reflecting their risk preference, endowing them with confidence

[50] These treatments were conducted on Prolific with 130 subjects per treatment. Details about experimental design are presented in Online Appendix B.

[51] This has been shown to be challenging for many subjects; see Esponda & Vespa (2014), Ali, Mihm, Siga & Tergiman (2021).

in their suboptimal choice.

Our baseline treatment *Primitives (Voting)* corresponds to exactly this case. As in our original experiment, subjects submit initial responses unaware the the task will be repeated. After submitting the first answer, they are asked (unincentivized) about their confidence in their initial answer using a 1-5 scale slider.[52]

Subsequently, we repeat the task for a total of 99 rounds. In between rounds, subjects receive information indicative of optimal behavior. We provided feedback with the same characteristics as in our original treatments, that is, feedback corresponds to natural sampling and is independent of subjects' choices. Specifically, in odd (even) rounds subjects learn the payoff of a random participant who voted for option 1 (option 2).[53] Learning is particularly easy here since there is a dominant action: Voting for option 1 always generates a payment of \$6, while voting for option 2 generates a payment of \$6 with 60 percent probability and \$10 with 40 percent probability. In particular, it is straightforward to notice that option 2 never pays \$0.

In *NoPrimitives (Voting)*, everything is identical to *Primitives (Voting)* except that, as in the comparison between our core treatments, we do not provide subjects with the numerical values of any of the primitives in the problem. Specifically, in the instructions, payments \$0, \$6 and \$10 are replaced by unknown variables A, B, C; in addition, subjects know that the computer knows the random number determining the payoff of option 2, but do not know whether or how the computer uses this information. Feedback is provided in the exact same way as in *Primitives (Voting)*. A comparison between *Primitives (Voting)* and *NoPrimitives (Voting)* provides a test that is similar in nature to the comparison between our core treatments (*Primitives* and *NoPrimitives*). Extrapolating from our earlier results, we expect that subjects in *Primitives (Voting)* will be relatively confident in their initial answer but that in the long run participants will make better choices in *NoPrimitives (Voting)* than in *Primitives (Voting)*.

Results are summarized in the top portion of Table 1. First, notice that mean and median first-round confidence in *Primitives (Voting)* is significantly higher relative to *NoPrimitives (Voting)* (p-value < 0.001 in both cases). However, the frequency of last-round optimal choices in *NoPrimitives (Voting)* is close to 75 percent and is significantly higher than the 57 percent of the *Primitives (Voting)* treatment (p-value 0.003). Approximately one-third of subjects responded optimally in

[52] Specifically, we ask them: 'How confident do you feel about your choice in Part 1?'

[53] If we provided payoff feedback directly on subjects' choices in this problem, a subject who votes for option 1 would not have the opportunity learn: they would just observe a payoff of \$6 in every round. In general, as pointed out in the introduction, feedback that is endogenous to the subject's choices can affect learning as has been shown in the literature (e.g. Esponda & Vespa (2018), Fudenberg & Vespa (2019)). In this paper, we abstract from this factor.

the first round of *Primitives (Voting)*, but if we focus on those who selected the suboptimal option 1 in the first round of both treatments, there is an even larger difference in long-run behavior. Approximately 70 percent of these subjects in *NoPrimitives (Voting)* are optimally voting for option 2 in the last round, but the number goes down to 43 percent in *Primitives (Voting)*.[54] These results are in line with the hypothesis that confidence in a suboptimal initial answer, driven by an incorrect understanding of the environment, results in lower levels of optimal behavior in the long run.

The other two treatments are generated to test the hypothesis that when subjects in an environment with primitives do not have as much confidence in their initial answer, they remain attentive to feedback. Thus, long-run behavior would not depend on whether primitives are initially provided or not. Specifically, *Complex Primitives (Voting)* involves the same problem as *Primitives (Voting)*, except that options are described deliberately in a more involved manner.[55] We hypothesized that subjects would be less confident in their initial answers in this treatment as the presentation makes the 'safe' vs. 'risky' framing not transparent. We also conduct a *Complex NoPrimitives (Voting)* treatment transforming the problem we just described in the same way as for *NoPrimitives (Voting)*. Feedback is provided in an identical manner in all four treatments.

Results for these treatments are summarized at the bottom of Table 1. We first point out that while there is a small but significant difference in average confidence, this is driven by a few outliers. In fact, median confidence in both treatments is the same and at the center of the scale. In terms of long-run choices, we now report no differences between treatments regardless of whether we focus on all subjects, or condition on whether subjects make an optimal round-one choice or not.[56] Note also that the rate of optimal last-round choices in *Complex Primitives (Voting)* is similar to that of *NoPrimitives (Voting)*. This evidence is consistent with the hypothesis that if subjects are less confident in an initial incorrect answer, they are more likely to learn in the long run.

Result #6: *Long-run behavior is more optimal in the voting problem when payoff-relevant prim-*

[54] Meanwhile the table also shows that there is essentially no last-round difference across treatments for subjects who selected optimally in round 1. For further analysis on these treatments see Online Appendix I.

[55] Option 1 is described as paying \$6 if there is only one vote for option 1; if there are two votes for option 1, it pays \$6 if the random number is smaller than or equal to 60, \$0 if the random number is between 61 and 70, \$10 if the random number is higher than 70. Notice that since option 1 can only have two votes when the computer votes for it, and the computer votes for it whenever the random number is lower than 60, option 1 will always pay \$6 as in *Primitives (Voting)*. Option 2 pays \$0 if the random number is smaller than or equal to 58, \$6 if the random number is 59 or 60, and \$10 otherwise. Notice that since option 2 is implemented if there are two votes for it and the computer votes for it whenever the random number is higher than 60, then voting for option 2 will either pay \$6 (when the computer votes for option 1) or \$10, as in *Primitives (Voting)*.

[56] The proportion of optimal choices in the last round of *Complex Primitives(Voting)* at 70 percent is significantly higher (p-value 0.029) than the 56.9 percent in *Primitives (Voting)*, despite evidence suggesting that learning in the Complex case is more challenging; see Online Appendix I.

Table 1: Optimality of Long-Run Behavior and Confidence in Voting

	Optimality of Vote in Last Round (in %)			Confidence	
	All	R1 Optimal	R1 Not Optimal	Mean	Median
Primitivites (Voting)	56.9	84.1	43.0	3.76	4.00
NoPrimitivites (Voting)	74.6	78.8	70.3	2.55	2.50
Δ	17.7	-5.3	27.3	-1.21	-1.5
p-value	0.003	0.493	0.001	0.001	<0.001
Complex Primitivites (Voting)	70.0	87.2	57.3	3.39	3.00
Complex NoPrimitives (Voting)	73.1	78.8	69.2	2.76	3.00
Δ	3.1	-8.4	11.9	-0.63	0.00
p-value	0.584	0.248	0.128	< 0.001	1.00

Note: To test for significance we use OLS. The left-hand side variable is the last-round choice (1=correct) in the first three columns of results. The sample in the second column of results is constrained to subjects who answered round 1 (R1) optimally, while the third on subjects who answer round 1 incorrectly. In the case of confidence, the right-hand side variable is the confidence measure where 5 is extremely confident and 1 indicates no confident at all. For the median we use quantal regressions.

itives are not provided. This replicates our main result (#1) in a new setting. Complicating the framing of the problem, and hence lowering confidence in initial answer, eliminates such a treatment effect.

6 Conclusion

We studied the persistence of mistakes in the presence of feedback and brought to light the different mechanisms that hinder learning from feedback. Our findings suggest mistakes are more likely to be persistent when they are driven by incorrect mental models that miss or misrepresent important aspects of the environment. Such models induce confidence in initial answers, limiting engagement with and learning from feedback. This insight also connects closely with the literature on learning with misspecified models and learning with endogenous attention, as we discussed in the introduction.

While it is beyond the scope of this paper to study persistence of every mistake in the presence of information, it is useful to think about the implications of our results for other biases. Our results suggest that learning from feedback might be easier in settings where agents make suboptiomal decisions but are aware of the fact that they are using mental shortcuts to avoid costs associated with identifying the optimal response, as in satisficing (Caplin, Dean & Martin 2011), but harder in settings where suboptimal behavior is driven by conceptual mistakes agents are less likely to be aware of, as documented here for base rate neglect and pivotal voting, but also likely with the winner's curse or the Monty Hall problem.[57] Confidence measures in initial responses can be useful

[57] See e.g. James, Friedman, Louie & O'Meara (2018) for difficulties with the Monty Hall problem and Kagel & Levin (2002) for the winner's curse. Relatedly, Danz, Vesterlund & Wilson (2022) study belief elicitation using a binarized-

in differentiating between mistakes to identify ones where subjects are more or less self aware of the suboptimality of their behavior. This brings a new perspective to an emerging research focusing on eliciting such measures.[58]

It is also worth highlighting the types of interventions that did and did not facilitate learning in our experiments. Simply providing information that is indicative of optimal behavior was not sufficient to counter systematic biases. Instead, it is important to be able to target agents' engagement with this information. The results also reveal several counterintuitive interventions that were effective in inducing optimal behavior in the long run. First, we find that withholding information that agents consider as payoff-relevant can increase attentiveness to feedback and foster learning. Second, we find that informing agents directly about the suboptimality of their actions increases engagement with feedback. Third, we find that complicating the framing of the problem lowers confidence in initial answer, fostering learning from feedback, consequently improving optimality of long-run behavior. While the controlled environment of the laboratory provides a natural starting point to study the interaction between biases and learning and possible interventions to facilitate learning, we believe that further work should examine these issues and the validity of our results in prominent field applications.

References

Agranov, M., Dasgupta, U. & Schotter, A. (2020), 'Trust me: Communication and competition in psychological games', *Working Paper* .

Ali, S. N., Mihm, M., Siga, L. & Tergiman, C. (2021), 'Adverse and advantageous selection in the laboratory', *American Economic Review* **111**(7), 2152–78.

Araujo, F. A., Wang, S. W. & Wilson, A. J. (2021), *American Economic Journal: Microeconomics* **13**(4), 1–22.

Barrett, G. F. & Donald, S. G. (2003), 'Consistent tests for stochastic dominance', *Econometrica* **71**(1), 71–104.

scoring rule and find that providing subjects with clear details on the incentives may actually trigger heuristics that can lead to deviations from truth telling. In other words, providing subjects with detailed information that they cannot properly process can lead to suboptimal choices relative to a baseline in which such detailed information is not provided. This manipulation is reminiscent of our distinction between *Primitives* and *NoPrimitives.*

[58] See Enke & Graeber (2023) for a cognitive uncertainty measure, and Enke, Graeber & Oprea (2023) for evidence on how confidence varies among some of the most well known biases in behavioral economics.

Barron, K., Huck, S. & Jehiel, P. (2019), 'Everyday econometricians: Selection neglect and overoptimism when learning from others', *Working Paper* .

Bayona, A., Brandts, J. & Vives, X. (2020), 'Information frictions and market power: A laboratory study', *Games and Economic Behavior* .

Bénabou, R. & Tirole, J. (2003), 'Intrinsic and extrinsic motivation', *The review of economic studies* **70**(3), 489–520.

Bénabou, R. & Tirole, J. (2016), 'Mindful economics: The production, consumption, and value of beliefs', *Journal of Economic Perspectives* **30**(3), 141–64.

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

Bohren, J. A. & Hauser, D. N. (2021), 'Learning with heterogeneous misspecified models: Characterization and robustness', *Econometrica* **89**(6), 3025–3077.

Bordalo, P., Gennaioli, N. & Shleifer, A. (2013), 'Salience and consumer choice', *Journal of Political Economy* **121**(5), 803–843.

Brunnermeier, M. K. & Parker, J. A. (2005), 'Optimal expectations', *American Economic Review* **95**(4), 1092–1118.

Caplin, A. & Dean, M. (2015), 'Revealed preference, rational inattention, and costly information acquisition', *American Economic Review* **105**(7), 2183–2203.

Caplin, A., Dean, M. & Martin, D. (2011), 'Search and satisficing', *American Economic Review* **101**(7), 2899–2922.

Cason, T. N. & Plott, C. R. (2014), 'Misconceptions and game form recognition: Challenges to theories of revealed preference and framing', *Journal of Political Economy* **122**(6), 1235–1270.

Charness, G., Oprea, R. & Yuksel, S. (2021), 'How do people choose between biased information sources? evidence from a laboratory experiment', *Journal of the European Economic Association* **19**(3), 1656–1691.

Cipriani, M. & Guarino, A. (2009), 'Herd behavior in financial markets: an experiment with financial market professionals', *Journal of the European Economic Association* **7**(1), 206–233.

Cooper, D. J. & Kagel, J. H. (2009), 'The role of context and team play in cross-game learning', *Journal of the European Economic Association* **7**(5), 1101–1139.

Cooper, D. J. & Van Huyck, J. (2018), 'Coordination and transfer', *Experimental Economics* **21**(3), 487–512.

Cosmides, L. & Tooby, J. (1996), 'Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty', *cognition* **58**(1), 1–73.

Dal Bó, E., Dal Bó, P. & Eyster, E. (2018), 'The demand for bad policy when voters underappreciate equilibrium effects', *The Review of Economic Studies* **85**(2), 964–998.

Danz, D., Vesterlund, L. & Wilson, A. J. (2022), 'Belief elicitation and behavioral incentive compatibility', *American Economic Review* **112**(9), 2851–2883.

Dekel, E., Fudenberg, D. & Levine, D. (2004), 'Learning to play bayesian games', *Games and Economic Behavior* **46**(2), 282–303.

Enke, B. (2020), 'What you see is all there is', *The Quarterly Journal of Economics* **135**(3), 1363–1398.

Enke, B. & Graeber, T. (2023), 'Cognitive uncertainty', *The Quarterly Journal of Economics* **138**(4), 2021–2067.

Enke, B., Graeber, T. & Oprea, R. (2023), 'Confidence, self-selection, and bias in the aggregate', *American Economic Review* **113**(7), 1933–1966.

Enke, B. & Zimmermann, F. (2019), 'Correlation neglect in belief formation', *The Review of Economic Studies* **86**(1), 313–332.

Esponda, I. & Pouzo, D. (2016), 'Berk–nash equilibrium: A framework for modeling agents with misspecified models', *Econometrica* **84**(3), 1093–1130.

Esponda, I. & Vespa, E. (2014), 'Hypothetical thinking and information extraction in the laboratory', *American Economic Journal: Microeconomics* **6**(4), 180–202.

Esponda, I. & Vespa, E. (2018), 'Endogenous sample selection: A laboratory study', *Quantitative Economics* **9**(1), 183–216.

Esponda, I. & Vespa, E. (2023), 'Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory', *Review of Economic Studies (forthcoming)* .

Esponda, I., Vespa, E. & Yuksel, S. (2023*a*), 'Mental models and transfer learning', *AEA Papers and Proceedings* **113**, 659–664.

Esponda, I., Vespa, E. & Yuksel, S. (2023*b*), 'Replication data for: "mental models and learning: The case of base-rate neglect"', *American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor] https://doi.org/10.3886/E194236* .

Eyster, E. & Weizsäcker, G. (2010), 'Correlation neglect in financial decision-making', *Working Paper* .

Falk, A. & Zimmermann, F. (2018), 'Information processing and commitment', *The Economic Journal* **128**(613), 1983–2002.

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

Fudenberg, D. & Lanzani, G. (forthcoming), 'Which misspecifications persist?', *Theoretical Economics* .

Fudenberg, D. & Peysakhovich, A. (2016), 'Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem', *ACM Transactions on Economics and Computation (TEAC)* **4**(4), 1–18.

Fudenberg, D., Romanyuk, G. & Strack, P. (2017), 'Active learning with a misspecified prior', *Theoretical Economics* **12**(3), 1155–1189.

Fudenberg, D. & Vespa, E. (2019), 'Learning theory and heterogeneous play in a signaling-game experiment', *American Economic Journal: Microeconomics* **11**(4), 186–215.

Gabaix, X. (2014), 'A sparsity-based model of bounded rationality', *The Quarterly Journal of Economics* **129**(4), 1661–1710.

Gagnon-Bartsch, T., Rabin, M. & Schwartzstein, J. (2021), 'Channeled attention and stable errors', *Working paper* .

Gennaioli, N. & Shleifer, A. (2010), 'What comes to mind', *The Quarterly journal of economics* **125**(4), 1399–1433.

Graeber, T. (2022), 'Inattentive inference', *Journal of the European Economic Association* .

Greiner, B. (2015), 'Subject pool recruitment procedures: organizing experiments with orsee', *Journal of the Economic Science Association* **1**(1), 114–125.

Grether, D. M. (1980), 'Bayes rule as a descriptive model: The representativeness heuristic', *The Quarterly journal of economics* **95**(3), 537–557.

Handel, B. & Schwartzstein, J. (2018), 'Frictions or mental gaps: what's behind the information we (don't) use and when do we care?', *Journal of Economic Perspectives* **32**(1), 155–78.

Hanna, R., Mullainathan, S. & Schwartzstein, J. (2014), 'Learning through noticing: Theory and evidence from a field experiment', *The Quarterly Journal of Economics* **129**(3), 1311–1353.

He, K. & Libgober, J. (2023), 'Evolutionarily stable (mis) specifications: Theory and applications', *arXiv preprint arXiv:2012.15007* .

Heidhues, P., Kőszegi, B. & Strack, P. (2018), 'Unrealistic expectations and misguided learning', *Econometrica* **86**(4), 1159–1214.

Huck, S., Jehiel, P. & Rutter, T. (2011), 'Feedback spillover and analogy-based expectations: A multi-game experiment', *Games and Economic Behavior* **71**(2), 351–365.

Huffman, D., Raymond, C. & Shvets, J. (2022), 'Persistent overconfidence and biased memory: Evidence from managers', *American Economic Review* **112**(10), 3141–3175.

James, D., Friedman, D., Louie, C. & O'Meara, T. (2018), 'Dissecting the monty hall anomaly', *Economic Inquiry* **56**(3), 1817–1826.

Kagel, J. H. (1995), 'Cross-game learning: Experimental evidence from first-price and english common value auctions', *Economics Letters* **49**(2), 163–170.

Kagel, J. & Levin, D. (2002), *Common value auctions and the winner's curse*, Princeton Univ Pr.

Kahneman, D. & Tversky, A. (1972), 'On prediction and judgement', *ORI Research Monograph* **12**(4).

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

Lima, S. L. (1984), 'Downy woodpecker foraging behavior: efficient sampling in simple stochastic environments', *Ecology* **65**(1), 166–174.

Louis, P. (2015), 'The barrel of apples game: Contingent thinking, learning from observed actions, and strategic heterogeneity', *Working paper* .

Martin, D. & Muñoz-Rodriguez, E. (2019), 'Misperceiving mechanisms: Imperfect perception and the failure to recognize dominant strategies', *Working paper* .

Martínez-Marquina, A., Niederle, M. & Vespa, E. (2019), 'Failures in contingent reasoning: The role of uncertainty', *American Economic Review* **109**(10), 3437–74.

Montiel Olea, J. L., Ortoleva, P., Pai, M. M. & Prat, A. (2022), 'Competing models', *The Quarterly Journal of Economics* **137**(4), 2419–2457.

Moser, J. (2019), 'Hypothetical thinking and the winner's curse: an experimental investigation', *Theory and Decision* **87**(1), 17–56.

Ngangoué, M. K. & Weizsäcker, G. (2021), 'Learning from unrealized versus realized prices', *American Economic Journal: Microeconomics* **13**(2), 174–201.

Ortoleva, P. (2012), 'Modeling the change of paradigm: Non-bayesian reactions to unexpected news', *American Economic Review* **102**(6), 2410–2436.

Schwartzstein, J. (2014), 'Selective attention and learning', *Journal of the European Economic Association* **12**(6), 1423–1452.

Sims, C. A. (2003), 'Implications of rational inattention', *Journal of Monetary Economics* **50**(3), 665–690.

Toussaert, S. (2017), 'Intention-based reciprocity and signaling of intentions', *Journal of Economic Behavior & Organization* **137**, 132–144.

Zimmermann, F. (2020), 'The dynamics of motivated beliefs', *American Economic Review* **110**(2), 337–61.