# Mental Models and Learning: The Case of Base-Rate Neglect[*]

Ignacio Esponda
UCSB

Emanuel Vespa
UCSD

Sevgi Yuksel
UCSB

September 30, 2022

**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.

[*] We would like to thank the editors, anonymous referees and numerous seminar participants for helpful comments. We are grateful for financial support from the UCSB Academic Senate. Esponda: iesponda@ucsb.edu. Vespa: evespa@ucsd.edu. Yuksel: sevgi.yuksel@ucsb.edu.

# 1 Introduction

Behavioral economics has accumulated a wealth of evidence documenting systematic biases in decision making. Some well-known examples include base-rate neglect (Kahneman & Tversky 1973; Bar-Hillel 1980), overconfidence (Moore & Healy 2008; Mobius, Niederle, Niehaus & Rosenblat 2022), the sunk-cost effect (Thaler 1980; Arkes & Blumer 1985), the law of small numbers (Rabin 2000), under-exploring (Schotter & Braunstein 1981; Cox & Oaxaca 2000), and correlation neglect (Eyster & Weizsäcker 2010; Enke & Zimmermann 2019). An important question is whether such biases are self-corrected in the presence of feedback. On the one hand, one might expect such biases to vanish with experience if agents are accumulating lots of 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]

The goal of this paper is to study optimality of long-run behavior in the presence of feedback, and bring to light the different mechanisms that hinder learning from feedback. We do so by designing a laboratory experiment with two crucial features. First, we consider a baseline treatment where subjects face a decision problem in which useful information about the problem generates biased behavior, possibly by inducing an incorrect understanding of the environment. Subjects face the same decision problem for 200 rounds and receive transparent feedback that is informative and simple (natural sampling). This allows us to study the evolution of behavior in this setting with ample 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 have to rely on 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.

Importantly, in our baseline treatment, information we provide to the subjects induces one of the most well-documented biases in the literature, base-rate neglect. Base-rate neglect (BRN) describes the tendency to underuse prior beliefs (the base rate) when updating in light of new information relative to the Bayesian benchmark. 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

[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.

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).

Our experimental design involves many repetitions of the updating problem described in the motivating example above (but presented using a more neutral framing). While BRN is not the only deviation from the Bayesian benchmark observed in the data, it is 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 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. 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 on 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. That is, 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

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

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 the *NoPrimitives* which is nine percentage points lower than the value in *Primitives*.[3]

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 this.

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 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 42 percent, 12 percentage points lower than the value in *Primitives*. This suggests that high confidence in the round one response in *Primitives* plays a critical role in inhibiting learning from feedback.

Second, we study whether differences in long-run behavior documented between *Primitives* and *NoPrimitives* can be fully explained by differences in confidence in initial response. Specifically, we investigate whether there are also differences between these treatments in terms of attentiveness to feedback. That is, 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] To provide an answer to this question, as a first step, we conduct a set of diagnostic

[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

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 gives us a simple measure of engagement with feedback: the lock-in decision reveals 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 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 34 percent of subjects see all 200 rounds.

While these results establish that information on the primitives lowers willingness to engage with feedback, it remains an open question how much this impacts learning in *Primitives.* To help us answer this question, 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 attending to the feedback is effectively lowered to zero. Results show that when feedback is presented in this way, subjects are able to learn more in both cases. 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.

Next, we study further exactly what subjects are learning from feedback in *Primitives* when behavior moves close to the Bayesian benchmark. With this aim, we first introduce a series of manipulations that move almost all subjects very close to the Bayesian benchmark. Specifically, we provide subjects in *Primitives* and *NoPrimitives* with the empirical frequencies (for the 200 rounds they have observed, as well as for a larger data set, organized as a summary table). For subjects in *Primitives*, who were initially far from the Bayesian benchmark but ultimately adjust

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.

their behavior, are they simply responding to the data, or are they also learning how to properly use the primitives? To study this, we include 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. Indeed, we find evidence that learning is partially transferable to this new setting. A non-negligible amount of base-rate neglect remains in *Primitives*, but a much higher proportion appears in *NoPrimitives*. Effectively, in this round with new primitives the treatment effect reverses: average beliefs in *Primitives* are closer to the Bayesian benchmark than 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. 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. In other words, our results suggest misconceptions in *Primitives* endow subjects with unjustified confidence in their initial answers, but not in *NoPrimitives*.[5]

The learning model (presented in Section 4.2) is also useful in that it provides a formal framework to think about how initial misconceptions, or incorrect mental models, can impact learning from feedback. An initial misconception can be modelled simply as a strong prior that deviates from the true value. We use the term 'confidence' to denote the strength of this prior. We also conjecture that agents with strong priors might not exert as much effort to extract information from the feedback presented to them. We are using 'attentiveness' in this general way to encompass any frictions in how the agents acquire, process, and record information about past outcomes. Formally, all we mean is that less attentive agents observe noisier signals about optimal behavior.

Finally, to provide evidence on the generalizability of our results beyond the updating problem, we conduct four more treatments in a new 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

[5] It is possible that subjects in both *Primitives* and *NoPrimitives* are prone to neglect the base rate. But, only in the former case subjects have the information available to apply this incorrect perspective. More generally, while subjects in *NoPrimitives* can also form an incorrect initial understanding of the environment, our results show they are attentive and responsive to the feedback. Their behavior suggests that they are self-aware of the possibility of being quite far from optimal behavior in round one.

(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. We hypothesize that the complex description makes it more likely that subjects are aware of the possibility of a mistake in their initial responses. Consistent with our earlier findings on the role of confidence, 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.

These results have implications for how policies should be designed to counteract behavioral biases. For instance, our findings challenge the perspective that biases documented in initial responses are self-corrected with experience based on informative feedback. Instead, our results suggest that mistakes can be persistent even in information-rich environments where optimal behavior is easy to identify. Hence, to successfully mitigate these mistakes, in addition to providing agents with information, policy interventions would need to influence how agents engage with this information. For example, our results suggest that lowering cost of engaging with the feedback, even when this is information that could have been extracted at a small cost, can have a large impact on shifting behavior towards optimal choices. Our results also reveal some counterintuitive interventions that can be effective in inducing optimal behavior in the long run. An interesting implication of our results is that withholding payoff relevant information (as in *NoPrimitives*) can lead to long-run choices that are closer to optimal in contexts where such information is likely to induce misconceptions. Alternatively, informing people directly about the suboptimality of their actions (as we do in the updating problem) or increasing the complexity of the setting (as we do in the voting problem) decreases confidence and increases engagement with feedback, facilitating learning from feedback. In terms of the specific bias that we study, our findings suggest that more attention should be paid to theories that allow for agents who exhibit partial base-rate neglect (for a recent example, see Benjamin, Bodoh-Creed & Rabin 2019).

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

**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 as we outline below.

First, our results provide support for a growing literature in economics that studies the implications of incorrect, misspecified mental 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. Some examples include models of behavioral agents as developed by Gennaioli & Shleifer (2010), Bordalo, Gennaioli & Shleifer (2013), and Gabaix (2014). A related literature also 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.

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 (in the tradition of misspecified models), 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]

[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 (2021), 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

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

Our paper also relates to a literature that studies long-run outcomes in the presence of feedback, often in environments where well-known biases play a role. In many of these cases, it is challenging to identify the different 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 (e.g. Huck, Jehiel & Rutter 2011), or more importantly, endogenous to the subject's choices. Learning can also be cognitively challenging if agents face a dataset with sample selection (e.g. 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 (given her model of the world) makes choices such that the collected information does not challenge her understanding 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,

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.

[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.

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. The paradigm in this literature is to study the *description-experience gap* which compares accuracy of beliefs when subjects are only given the primitives to when subjects only have experience to rely on. For example, Gigerenzer & Hoffrage (1995) show that base-rate neglect is attenuated when subjects are provided with natural frequencies (instead of the underlying primitives). 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]

## 2 Experimental design

### 2.1 Main treatments and procedures

Our experimental design consists of two core and nine supporting treatments. The core treatments consist of four parts. The first two parts of the core treatments test the central hypothesis in the paper.[15] All other parts and treatments are designed to study the mechanisms underlying these results and the generalizability of these results to other settings. 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. The remaining two parts of the core treatments and the nine supporting treatments are introduced in Sections 4 and 5.

[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, 2021), 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.

[15] For expositional purposes we describe our core treatments in four parts, though the presentation for subjects was broken up further into nine parts. A full description of the experimental design for all treatments is provided in Online Appendix B.

### 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 on a binary state using a binary signal. 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. Subjects are told in both treatments that there are 100 projects, each either a success or a failure, and the task consists of assessing the chance that a randomly selected project is a success vs. a failure conditional on a signal that is informative about the type of the project. In *Primitives*, subjects know that 15 projects are successes and 85 projects are failures. 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. We frame the signal as the computer running a test on the selected project. The signal is either positive or negative. In *Primitives*, subjects also know that the signal has a reliability of 80 percent.[16] In *NoPrimitives*, subjects are told that the signal has a reliability of $q$ percent, but while we describe the meaning of $q$ just as in *Primitives*, we do not reveal the value of $q$. This parameterization (prior = .15, reliability of signal = .8) is the same for both treatments and corresponds to the classic parameterization of Kahneman & Tversky (1972).

To summarize, the *only* difference between the two treatments is that subjects know the prior and the reliability of the signal in *Primitives*, while these values are not provided to the subjects in *NoPrimitives*. All other parts of the instructions, in this part and in all subsequent parts, are identical. 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 test being positive ($B_{Pos}$), and (2) the belief that the project is a success conditional on the test being negative ($B_{Neg}$). In this round and in all future belief-elicitation rounds, subjects are incentivized using a standard incentive-compatible mechanism.[17]

In *Primitives*, subjects could in principle use Bayes' rule to provide the correct answer, but consistent with the literature we find substantial deviations from the Bayesian Benchmark, with

[16] The notion of reliability is carefully explained. Specifically, subjects are told that if the project is a success (failure), the test result will be positive (negative) with 80 percent chance and negative (positive) with 20 percent chance.

[17] 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.

BRN being the dominant mistake (see Section 3.1). 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.

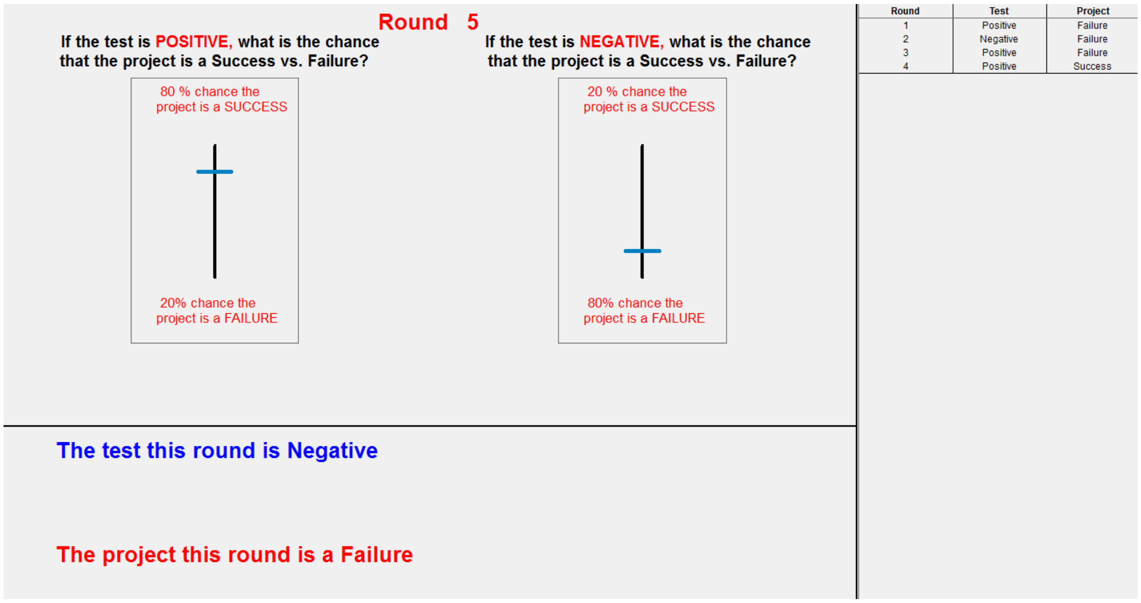
### II. Learning: Repetition of updating task, rounds 2-200

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.[18] 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 (test result 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 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. For example, a subject who just submitted responses for round 110 would see the outcomes for each of rounds 110 through 119 without being asked again for her beliefs until round 120; the same procedure follows in blocks of ten rounds. 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.

Importantly, the instructions in both treatments stress that subjects will be facing the exact same environment in every round. That is, the reliability of the signal and the prior are the same in all rounds, but the state is drawn independently and with replacement in every round.

[18] Each part is introduced as a surprise, meaning that subjects were not informed in advance of what latter 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

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

The test this round is Negative

The project this round is a Failure

Figure 1: Interface Screenshot at Round Five of Core Treatments

### 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. We provide more details about the experimental procedures 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.

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). The experiment, which lasted 90 minutes, was conducted using zTree (Fischbacher 2007). In addition 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.

[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.

### 2.2 Bayesian and Base-rate neglect benchmarks

Given the prior $p = .15$ (ex-ante probability that the project is a success) 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 $B^{Bay}_{Neg} = 4\%$. Let $(B_{Pos}, B_{Neg})$ capture the subject's reported beliefs in percentage terms, namely, their assessment that the project is a success conditional on a positive and negative signal, respectively. Throughout the paper, we refer to a subject's beliefs as Bayesian (in a given round/part) if $(B_{Pos}, B_{Neg}) = (B^{Bay}_{Pos}, B^{Bay}_{Neg})$.[20]

A perfect Base Rate Neglect (pBRN) response fully ignores the prior (treating it as uniform), so that in percentage points we have: $(B^{pBRN}_{Pos}, B^{pBRN}_{Neg}) = (80, 20)$. Thus, particularly relevant for our *Primitives* treatment, we refer to a subject's beliefs as being consistent with pBRN if $(B_{Pos}, B_{Neg}) = (B^{pBRN}_{Pos}, B^{pBRN}_{Neg})$.

### 2.3 Discussion of the core treatments

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. Responses in round one, where the main task is first introduced, provide a benchmark for beliefs in the absence of feedback. Long-run beliefs in later parts allow us to observe the impact of feedback. Feedback is natural (the state-signal realization of that round), 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.

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

We organize our main results as follows: In Section 3.1, we confirm that initial (i.e., round one) responses in *Primitives* replicate previous findings in the literature related to BRN. In Section

[20] Subjects in *Primitives* could, in principle, submit Bayesian posteriors in all rounds. In *NoPrimitives*, this was clearly not possible in every round, but only approximately possible (by learning from feedback) in the long run.

3.2, focusing on evolution of beliefs with 200 rounds of feedback, we 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 *Primitives*

Here, we summarize patterns in initial responses in *Primitives*, that is, round one of the updating task. 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).[21] 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 computation errors by the subjects.[22] 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.

In the upcoming sections, we will present more details on the distribution of beliefs in *Primitives* and contrast it to *NoPrimitives*. The main message from this section is 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). Next, we study choices in rounds 1-200 at the aggregate level to evaluate to what extent feedback can correct such behavior.

### 3.2 Learning in *Primitives* vs. *NoPrimitives*

We start by describing evolution of beliefs at the aggregate-level across rounds in *Primitives* vs. *NoPrimitives* to evaluate consistency of beliefs with the Bayesian Benchmark at different feedback levels.[23] Figure 2 presents the evolution of beliefs in *Primitives* using red squares–where the number next to a square indicates the round that the average corresponds to. The average round one beliefs

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

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

[23] On average, subjects will experience 29 (58) rounds with a positive and 71 (142) rounds with a negative signal by the end of 100 rounds (200 rounds).

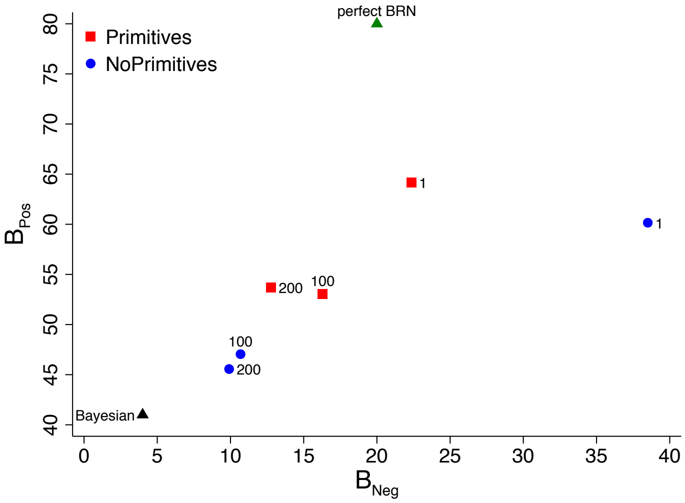


Figure 2: Evolution of average beliefs

Notes: The vertical (horizontal) axis represents beliefs conditional on the signal being positive (negative). Triangles indicate the Bayesian and pBRN benchmarks. Squares (Circles) report averages in *Primitives* (*NoPrimitives*). The numbers indicate the round for which the averages are reported.

in *Primitives* are $(B_{Pos}, B_{Neg}) = (64, 22)$. We observe that average beliefs in this treatment move closer to the Bayesian benchmark (and away from the pBRN point) with experience: After 100 rounds, average beliefs are $(53, 16)$, which corresponds to an adjustment towards the Bayesian benchmark of about eleven and six percentage points in $B_{Pos}$ and $B_{Neg}$, respectively. However, at this point, average beliefs are still twelve percentage points away from the Bayesian benchmark conditional on either signal.

The evidence suggests that while beliefs move towards the Bayesian benchmark with experience, the adjustment is slow and partial after 100 rounds. Note, however that there could be many factors that slow down learning in such a setting. 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 $(47, 11)$. Figure 2 indicates that after 100 rounds there is a treatment effect of approximately six percentage points in each dimension. That is, a first look at the evidence suggests that learning from feedback is more difficult in *Primitives*. The figure reveals a similar conclusion if we look at rounds 101-200.

The evolution of average beliefs across all rounds is presented in Figure 3. Shaded lines in the background show average beliefs in each round of each treatment. Darker lines in the foreground,

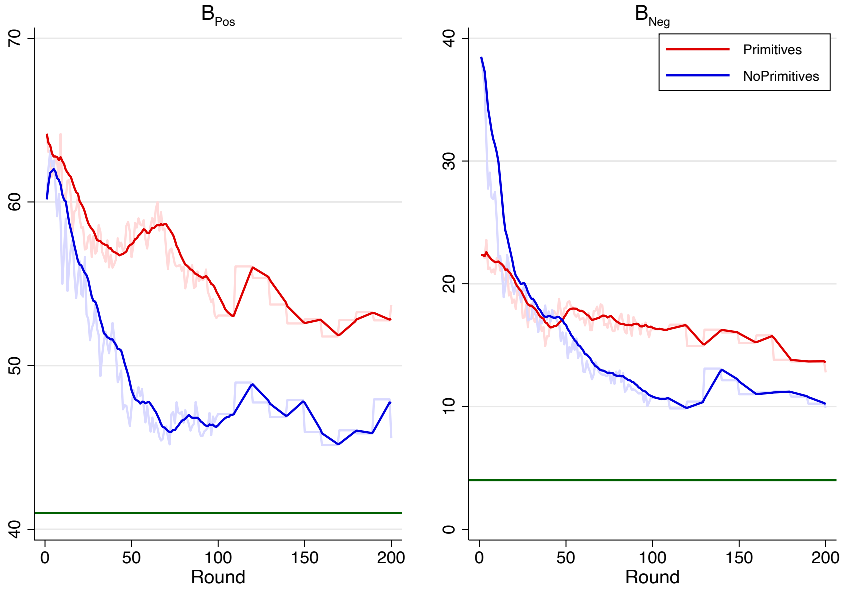


Figure 3: Evolution of beliefs

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.

which show beliefs averaged over prior ten rounds, provide a smoother depiction of the evolution of beliefs and make differences between the treatments more discernible. The main results can be seen in Figure 3 as well. First, 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. In addition, the figure reveals that in both treatments adjustments mostly take place during the first 100 rounds.

To provide statistical analysis on the differences between *Primitives* and *NoPrimitives* depicted in Figures 2 and 3, we focus on two questions: (1) Are there treatment differences in how far beliefs are to the Bayesian benchmark? (2) Are beliefs different between the two treatments?[24,25]

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

[24] 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).

[25] 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.

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. Since we have two observations per subject per round ($B_{Pos}$ and $B_{Neg}$), we estimate both regressions (on $B_{Pos}$ and $B_{Neg}$) as a system using seemingly unrelated regressions, allowing errors to be correlated for a fixed subject (but independent across subjects). Estimating the regressions as a system of equations allows us to test the null hypothesis of no treatment effect.[26]

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).

We next summarize our results on long-run beliefs in the presence of feedback:

***Result #1:*** *Beliefs in both treatments move closer to the Bayesian benchmark from round 1 to 200. By round 200, beliefs in NoPrimitives are significantly different from beliefs in Primitives, and beliefs in NoPrimitives are significantly closer to the Bayesian benchmark than beliefs in Primitives.*

## 3.3 Heterogeneity

To provide an overview of the heterogeneity in responses, Figures 4 and 5 present the distribution of beliefs in *Primitives* and *NoPrimitives* at different feedback levels. As can be seen in the left plot of Figure 4, 56.3 percent of subjects in round one of treatment *Primitives* submit beliefs that are consistent with pBRN. There are a few other points around which beliefs are somewhat concentrated. For instance, 4.7 percent of subjects have Bayesian beliefs, and 7.8 percent of subject display signal neglect (i.e. beliefs conditional on either signal are equal to the prior). The right plot of Figure 4 shows that by round 200 the distribution of beliefs has shifted significantly and that most subjects can essentially be categorized into two groups. There is a large cluster close to or at the pBRN point and another close to or at the Bayesian point.[27]

[26] Specifically, at a given round, for each possible signal $j \in \{Neg, Pos\}$, we estimate the following system of equations using seemingly unrelated regressions: $b_j = \alpha_j + \beta_j P + \upsilon_j$, where $b_j$ corresponds to either $|B_j - B_j^{Bay}|$ or $B_j$ depending on which question we are answering; $\upsilon_j$ is an error term; and $P$ is a dummy that takes value 1 if the variable comes from *Primitives.* The p-values that we report to evaluate treatment effects result from using a Wald test on the hypothesis that both treatment coefficient estimates are equal to zero (i.e. $\beta_{Neg} = \beta_{Pos} = 0$). Tables 4 and 5 in Online Appendix C.1 also reports results separately conditional on a positive and negative signal.

[27] Thirty-five percent of subjects are at ±10 percentage points of the pBRN benchmark and the similar proportion is within ±10 percentage points of the realized frequencies.

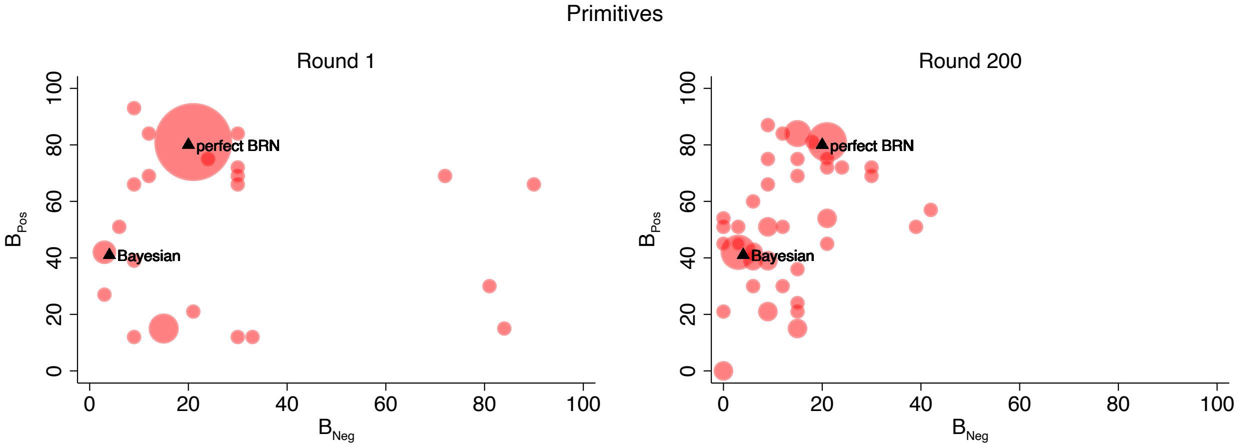


Figure 4: Density plots for *Primitives*

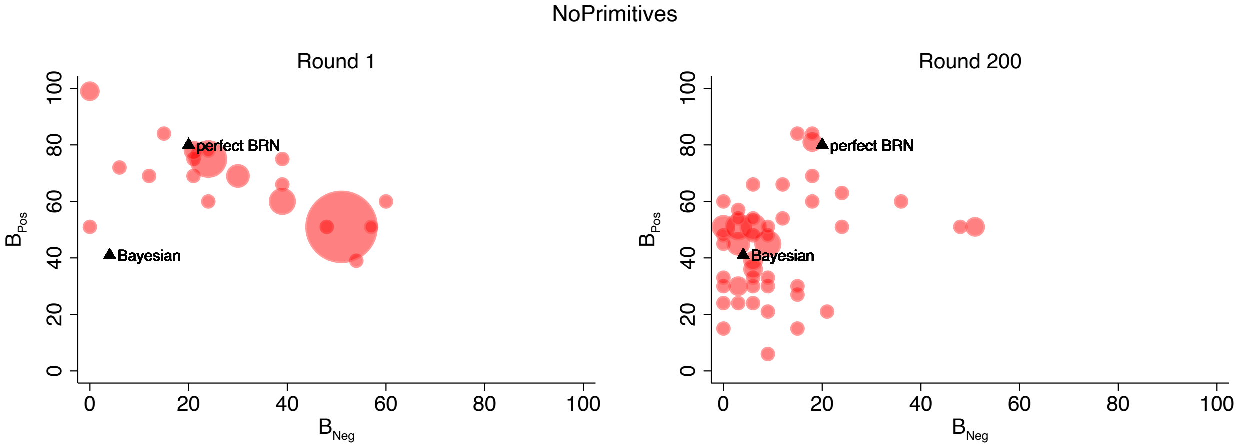


Figure 5: Density plots for *NoPrimitives*

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.

As can be seen in the left plot of Figure 5, subjects' beliefs in round one of *NoPrimitives* are quite different from those of *Primitives*, though they are approximately equidistant to the Bayesian benchmark. These beliefs can largely be organized into two groups. A large mass of subjects (forty-five percent) submit $(B_{Pos}, B_{Neg}) = (50, 50)$. 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 (fifty-two percent) submit beliefs that suggest they consider the labels we used for the signals (positive vs. negative) to provide some information. That is, their beliefs indicate that a randomly selected project is more likely to be a success conditional on a positive vs. a negative signal ($B_{Pos} > B_{Neg}$). By round 200 (right plot of Figure 5), the mass at $(50, 50)$ largely disappears and fifty-two percent of subjects are at ±10 percentage points of the realized frequencies. We summarize these observations:

***Observation:*** *In Primitives, beliefs for the 56.3 percent of the subjects are consistent with pBRN in round one, but are split into two equally-sized groups by round 200: those close to the pBRN benchmark and those close to the Bayesian benchmark. In NoPrimitives, subjects' beliefs in round one are either* $(50, 50)$ *or ranked to reflect informativeness of labels (*$B_{Pos} > B_{Neg}$*), and by round 200, beliefs for 52 percent of subjects are close to the Bayesian benchmark.*

These patterns suggest long-run differences between *Primitives* and *NoPrimitives* to be driven by those subjects who initially display perfect base-rate neglect in *Primitives.* In Online Appendix C.1, we study this more closely by contrasting long-run beliefs of those subjects in *Primitives* who initially display perfect base-rate to others in the same treatment. We confirm that there is a significant treatment effect (relative to *NoPrimitives*) only for the first group. Other subjects in *Primitives* have beliefs that are on average similar to the beliefs of subjects in *NoPrimitives* by round 200.[28]

## 4 Mechanisms

In the previous section we established that long-run beliefs are farther away from the Bayesian benchmark in a treatment where subjects were given information on the primitives of the problem (which in principle enables them to identify optimal behavior from round one) relative to a treatment where such information was withheld from the subjects. In this section, we study how this is possible by identifying the main mechanisms underlying this result.

We begin by providing a discussion of plausible mechanisms. 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 their round one answer. This type of incorrect understanding might make subjects in *Primitives* more confident in their initial response. Here, we use the term "confidence" to capture how strong the agent's prior

[28] In *Primitives* subjects whose round one responses are consistent with pBRN end up with average beliefs $B_{pos} = 61$ and $B_{neg} = 14$. The corresponding beliefs for others in same treatment are 45 and 11, respectively. For comparison average belief in round 200 of *NoPrimitives* is $B_{pos} = 46$ and $B_{neg} = 10$.

beliefs are about the optimality of their responses in round one. Note that 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.

The second mechanism is closely tied to the first as it 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. 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. It could be that learning in *Primitives* is more difficult because subjects are less attentive (in a general sense) to the feedback. See Section 4.2 for further discussion of what we mean to capture by *attentiveness*.

In this section, we provide evidence on the extent to which the mechanisms discussed above play a role in hindering learning from feedback. In Section 4.1, we introduce a new treatment and establish that confidence plays an important role in slowing down learning. In Section 4.2, we introduce two separate treatments and find that attentiveness also plays an important role. Finally, in Section 4.3, we use data from two additional treatments (and the core treatments) to structurally estimate a learning model, which enables us to assess the relative importance of confidence and attentiveness in our core treatments. We estimate that, by round 200, roughly half the treatment effect we observe between *Primitives* and *NoPrimitives* can be attributed to confidence alone (as measured by the strength of the prior) and the other half to attentiveness (as measured by the precision of information contained in the feedback).

## 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.[29] This

[29] 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. Subjects who submitted a Bayesian response to both questions didn't receive any message.

treatment includes 70 subjects and was conducted at UCSD.[30]

Since the two treatments are identical up to end of round one, we begin by confirming that the distribution of responses in round one are similar between these treatments. In particular, the distribution of responses (for $B_{Pos}$ or $B_{Neg}$) are not significantly different between *Primitives* and *Primitives w/ shock* according to a Kolmogorov-Smirnov test (p-value 0.451 for $B_{Pos}$ and 0.968 for $B_{Pos}$). More specifically, in *Primitives w/ shock*, average round one response is 59 (as opposed to 64 in *Primitives*) for $B_{Pos}$ and 25 (as opposed to 23 in *Primitives*) for $B_{Neg}$, and we cannot reject the null hypothesis that beliefs are jointly (for $B_{Pos}$ and $B_{Neg}$) the same in both treatments (p-value 0.419). In addition, the two treatments are not different in terms of distance to the Bayesian benchmark (p-value 0.159). The qualitative differences in average beliefs between the two treatments are driven by slightly lower share of pBRN subjects: 49 percent in *Primitives w/ shock* as opposed to 56 percent in *Primitives* (although, these shares are not statistically different, p-value 0.374).

Having established that round one behavior is similar between *Primitives* and *Primitives w/ shock*, we can study the effect of the message inducing a confidence shock on long-run behavior by contrasting beliefs in these two treatments. Given round one responses, 90 percent of subjects in *Primitives w/ shock* received a message that stated both of their initial answers (on $B_{Pos}$ or $B_{Neg}$) 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.

In Online Appendix D, we study heterogeneity in the data in *Primitives w/shock* by replicating

[30] We conducted some treatments at a different UC campus due to subject pool constraints and timing differences in when experimental laboratories were reopened during the Covid pandemic.

[31] In addition, three percent of subjects received a message indicating that at least one of their answers were incorrect.

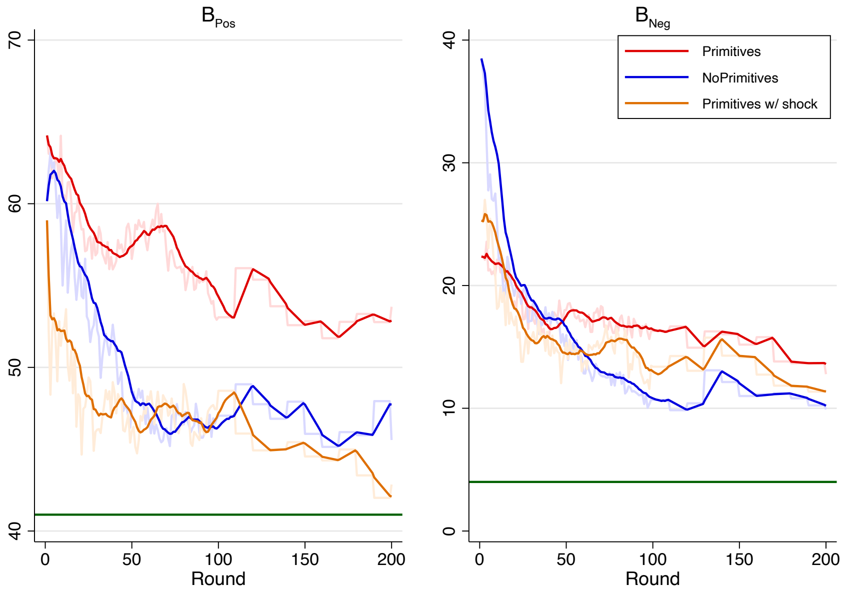


Figure 6: Comparing Evolution of Beliefs in *Primitives w/ shock* to *Primitives* and *NoPrimitives*

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.

the density plots of Figure 4 for this treatment. As expected, the density plot for round one closely resembles the one for *Primitives*. However, by round 200, the density plot for *Primitives w/shock* looks different from the one for *Primitives* and much closer to the one for *NoPrimitives*. The density plot reveals that the majority of responses in round 200 of *Primitives w/shock* are clustered around the Bayesian Benchmark. These differences further demonstrate the effectiveness of the message in this treatment in facilitating learning from feedback.

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. 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.

***Result #2:*** *Shocking confidence of subjects in their initial response (by telling them their* *answers are incorrect) improves optimality of long-run behavior. By round 200, beliefs in Primitives w/ shock are different, and closer to the Bayesian benchmark, relative to Primitives, and not different from those in NoPrimitives.*

## 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*.

It is useful at this point to discuss further what we mean by *attentiveness* and why it is difficult to study experimentally. First, the experiment was designed such that the feedback was visually available to the subjects at any point at almost no cost.[32] With attentiveness, we mean to capture a more meaningful notion in which subjects are not just looking at the data but are also engaging with it in a way that could be effective in changing their beliefs. Note that given the stochastic nature of the task no single round of feedback can invalidate a subject's beliefs. Learning from feedback requires subjects to process the feedback in a way that generates a compelling test of their model of the world. For example, looking at 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, there is still a processing cost associated with extracting the signal from the data. 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—and hence fail to fully grasp this pattern. This is precisely the type of inattentiveness we hope to capture in the new experiment.

Studying the degree to which learning is slowed down by partial attentiveness to the feedback is made difficult by the fact that it is not possible to directly observe attentiveness (as defined above) in our core treatments. To overcome this challenge, we run two diagnostic treatments. These treatments labeled as *Primitives w/ lock in* and *NoPrimitives w/ lock in*, are are identical, respectively,

[32] For example, the outcome of each round was prominently presented and the subjects were required to click on buttons on the same screen to proceed to the next round. Moreover, the outcome of each round was automatically recorded and displayed in a history table in all future rounds. While in principle subjects may actively try not to observe portions of their screens, these design features were put in place to minimize the costs associated with seeing and keeping track of the data.

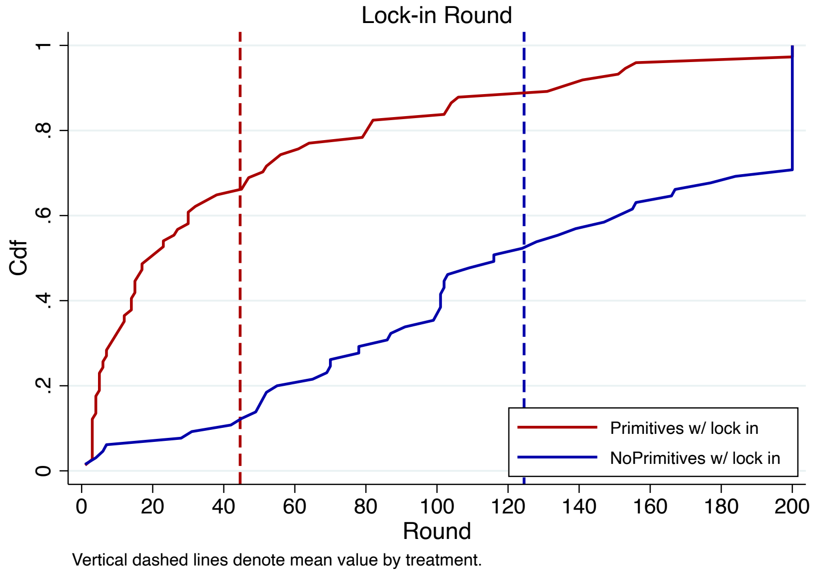


Figure 7: Evolution of beliefs

Notes: Subjects who never locked in are coded as locked in at round 200. Vertical lines denote mean values.

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. These treatments were run at UCSD and consist of 74 and 65 subjects in *Primitives w/ lock in* and *NoPrimitives w/ lock in*, respectively.[34]

Note that the new treatments are identical to our main treatments—*Primitives* and *NoPrimitives*—in round one. In Online Appendix E we confirm that initial responses are similar between the core treatments and the new ones with the lock in option.[35]

[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] Differences in number of subjects across the two treatments was not intentional and is driven by differences in show-up rates to the three sessions conducted per treatment.

[35] One difference is that there are slightly fewer pBRN subjects in *Primitives w/ lock in* relative to *Primitives*: 42 percent vs. 56 percent ($p = 0.094$). 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.

We focus on the main diagnostic purpose of these treatments: measuring engagement with feedback.[36] For this purpose, Figure 7 shows the cumulative distribution of round of lock in decisions in *Primitives w/ lock in* and *NoPrimitives w/ lock in*. 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*.[37] In *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 34 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 $p < 0.000$).

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*.

***Result #3:*** *Subjects lock in their choices earlier in Primitives w/ lock in relative to NoPrimitives w/ lock in. This points to differences in willingness to be attentive to feedback between our core treatments (Primivites vs. NoPrimitives).*

### 4.3 Quantifying the importance of attentiveness

We have established that confidence in an initially incorrect answer can negatively impact the optimality of long-run behavior. This can occur for two related reasons: Subjects place more weight on a stronger prior, and subjects are less attentive to feedback. We have also established that subjects are indeed less attentive to feedback, so at this point we would like to assess the

[36] Clearly, lock in decisions will impact learning from feedback. Since partial attentiveness in our main treatments most likely takes a very different form than stopping to observe feedback altogether after a certain round (which is what happens in the lock in treatments), we do not attempt to draw connections between learning dynamics in these new treatments and our main treatments. Nonetheless, Figure 21 in Online Appendix E replicates Figure 3 for these new treatments.

[37] We test for first-order stochastic dominance using the test in Barrett & Donald (2003). The test consists of two 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 <0.000). 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 0.829).

relative importance of prior strength and attentiveness, since these mechanisms have different policy implications regarding how to correct biases.

In particular, our objective in this section is to assess 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 would the gap in distance to Bayesian benchmark (as measured in round 200) 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 model of learning and new treatments.

### A model of learning

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 $\eta$, such that:

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

While $p_0$ denotes the expected value of $p$, $\eta$ captures the strength of the prior and, hence, can be interpreted as a measure of the agent's confidence.[38]

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 $\sigma n$ for some $\sigma \in [0,1]$ (but remembers $f$ correctly).[39] The agent's updated posterior is still characterized by a Beta distribution with adjusted parameters $\tilde{p}$ and $\tilde{\eta}$:

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

In summary, the model describes how beliefs evolve with feedback as a function of three parameters: $p_0$, prior expected value on $p$; $\eta$, a measure of initial confidence; and $\sigma$, attentiveness to

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

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

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 $\sigma/\eta$, it is not possible to separately identify these parameters.[40] To overcome this challenge, we run two new treatments, which we describe below. The goal of these treatments is to create environments in which prior beliefs (characterized by $p_0$ and $\eta$) are the same as in our core treatments, but where the cost of attentiveness is lowered to zero, such that $\sigma$ can reasonably be assumed to be 1.

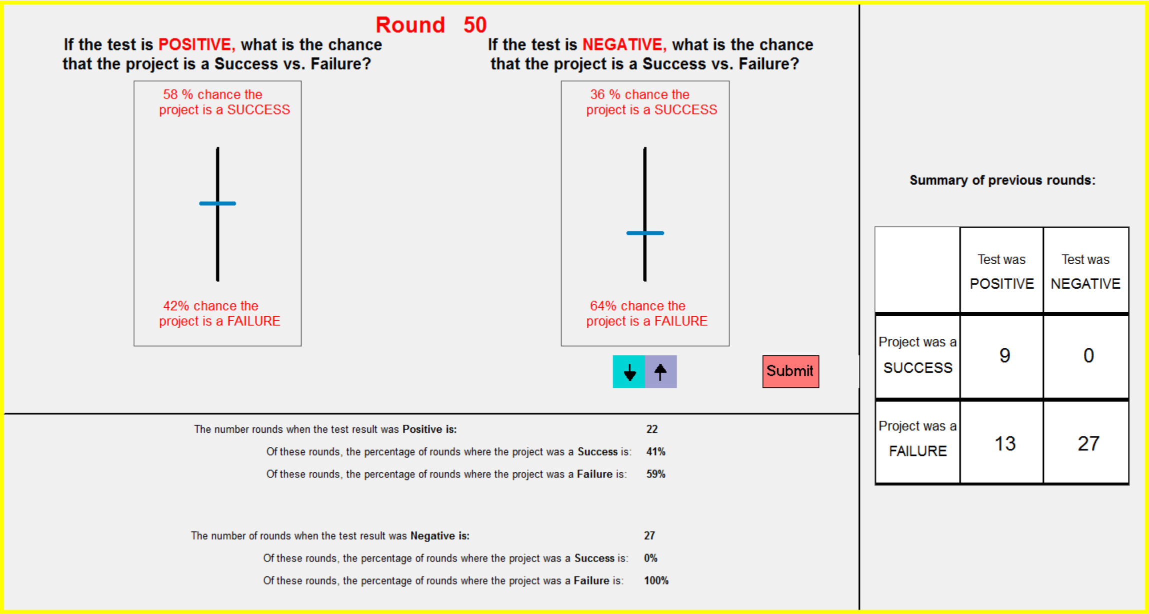
### New treatments where cost of attentiveness is minimized

We run 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). Critically, the feedback was not aggregated or processed. In *Primitives w/ freq* and *NoPrimitives w/ freq*, we still provide feedback on a round-by-round basis. But, in addition, feedback from all previous rounds is 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. The screenshot in Figure 8 shows a subject making a choice in round 50.

As mentioned earlier, the goal of these new treatments is to minimize the cost of attentiveness to feedback. The computer interface processes feedback in a natural way (computing empirical frequencies) and presents it to subjects in an easy-to-read format. To ensure that subjects are indeed aware of all this information presented to them, the interface also requires subjects to give us back this information (which is presented on the same screen) every 20 rounds.[41] Overall, we interpret these treatments as corresponding to an environment, where by design, the attentiveness parameter $\sigma$ is set to 1. These treatments were conducted at UCSB and consist of 59 subjects both in *Primitives w/ freq in* and *NoPrimitives w/ freq*, respectively.

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

[41] For details see Online Appendix B.



Round 50

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

58 % chance the project is a SUCCESS

42% chance the project is a FAILURE

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

36 % chance the project is a SUCCESS

64% chance the project is a FAILURE

The number rounds when the test result was **Positive is:** 22

Of these rounds, the percentage of rounds where the project was a **Success** is: 41%

Of these rounds, the percentage of rounds where the project was a **Failure** is: 59%

The number of rounds when the test result was **Negative is:** 27

Of these rounds, the percentage of rounds where the project was a **Success** is: 0%

Of these rounds, the percentage of rounds where the project was a **Failure** is: 100%

Summary of previous rounds:

| | Test was POSITIVE | Test was NEGATIVE |
|---|---|---|
| Project was a SUCCESS | 9 | 0 |
| Project was a FAILURE | 13 | 27 |

Figure 8: Interface Screenshot of Treatments with Frequencies

### Learning in the treatments with frequency information

Figure 9 depicts the evolution of beliefs with feedback in the treatments with frequency information and contrasts these to the core treatments.[42] The figure reveals stark differences in learning when feedback is presented in an aggregated form. For instance, in *Primitives w/ freq*, by round 200, $B_{Pos}$ is 10 percentage points lower (p-value 0.006) and $B_{Neg}$ is 5 percentage points lower (p-value 0.031) than those in *Primitives*. Beliefs ($B_{Pos}$ and $B_{Neg}$) are jointly different between between the core treatments and the ones with frequency information by round 200 (p-value is 0.010 between *Primitives w/ freq* and *Primitives*, and 0.033 between *NoPrimitives w/ freq* and *NoPrimitives*). Furthermore, providing feedback in a processed way impacts distance to the Bayesian benchmark. By round 200, beliefs are closer to the Bayesian benchmark in the treatments with frequency information relative beliefs in the corresponding core treatments (p-value $< 0.000$ when comparing *Primitives w/ freq* to *Primitives*, as well as *NoPrimitives w/ freq* to *NoPrimitives*).

The evidence also points towards convergence in behavior between the treatments with fre-

[42] 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. In *Primitives w/ freq*, average round one response is 67 (as opposed to 64 in *Primitives*) for $B_{Pos}$ and 22 (as opposed to 23 in *Primitives*) for $B_{Neg}$. These values are not jointly significantly different from those in *Primitives* (p-value 0.710).

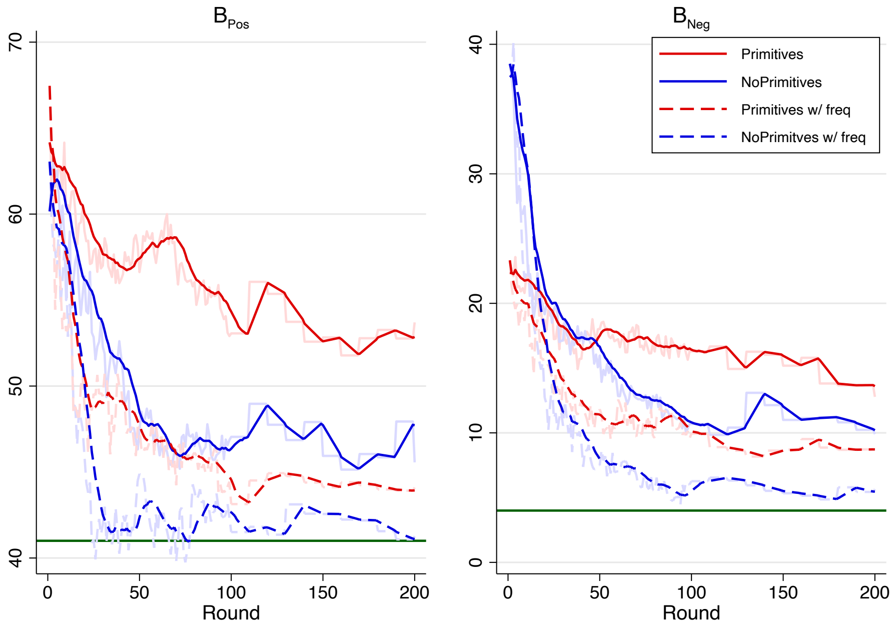


Figure 9: Evolution of Beliefs in Treatments with Frequencies Relative to Core Treatments

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.

quency information. While Figure 9 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 not observed in the new treatments (between *Primitives w/ freq* and *NoPrimitives w/ freq*). By round 200, beliefs are not different between *Primitives w/ freq* and *NoPrimitives w/ freq* (p-value 0.196), and are similar with respect to distance to Bayesian benchmark (p-value 0.313).[43] In Online Appendix F we replicate the density plots of Figures 4 and 5 (for round one and 200) for these new treatments. The figures reinforce the results reported here: Distributions look very similar to the core treatments in round one, but there is a tighter and bigger cluster around the Bayesian Benchmark in round 200. The unequivocal improvements in the optimality of long-run behavior observed in the new treatments, *Primitives w/ freq* and *NoPrimitives w/ freq*, strongly suggest that there were non-negligible costs associated with attending to this

[43] We also cannot reject that beliefs are coming from the same distribution for $B_{Pos}$ or $B_{Neg}$ according to a KolmogorovSmirnov test (p-values 0.499 and 0.365, respectively).

feedback in our core treatments (where the feedback was readily available but was not aggregated and processed for subjects).

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.* However, these results don't establish whether subjects are differentially attentive to the data in *Primitives* vs. *NoPrimitives.* In the next section, we make use of the learning model to answer this question.

### Estimation and Counterfactual analysis

Our goal is to study the degree to which learning is hindered in our core treatments by (i) high confidence in initial response (as captured by $\eta$ in the model), and (ii) partial attentiveness to data (as captured by $\sigma$ in the model). Our strategy in this section is to make use of the treatments with frequency information to separately identify the strength of each channel. Specifically, we estimate $\eta$ from the new treatments taking $\sigma = 1$ in this setting. Then, taking as given the estimated values of $\eta$ (from the new treatments), we use data from the core treatments to estimate $\sigma$.

This exercise can be done in several different ways. Here, we present the simplest possible version where 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 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.

Figure 10 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 where feedback is presented in an aggregated and processed form to subjects (depicted using dashed lines in the figure), differences in speed of learning are attributed to differences in confidence. Specifically, our estimates for $\eta$ are substantially higher for those subjects who were given the primitives vs. those who were not (twice as high for $B_{Pos}$ and five times as high for $B_{Neg}$).[44] This confirms our initial

[44] Estimates of $\eta$ for $B_{Pos}$ are 4.2 and 2.2 in *Primitives* and *NoPrimitives*, respectively. The corresponding values are 5.9 and 25 for $B_{Neg}$. Statistical tests using bootstrapping show differences to be significant (p-value $< 0.000$ in both cases).

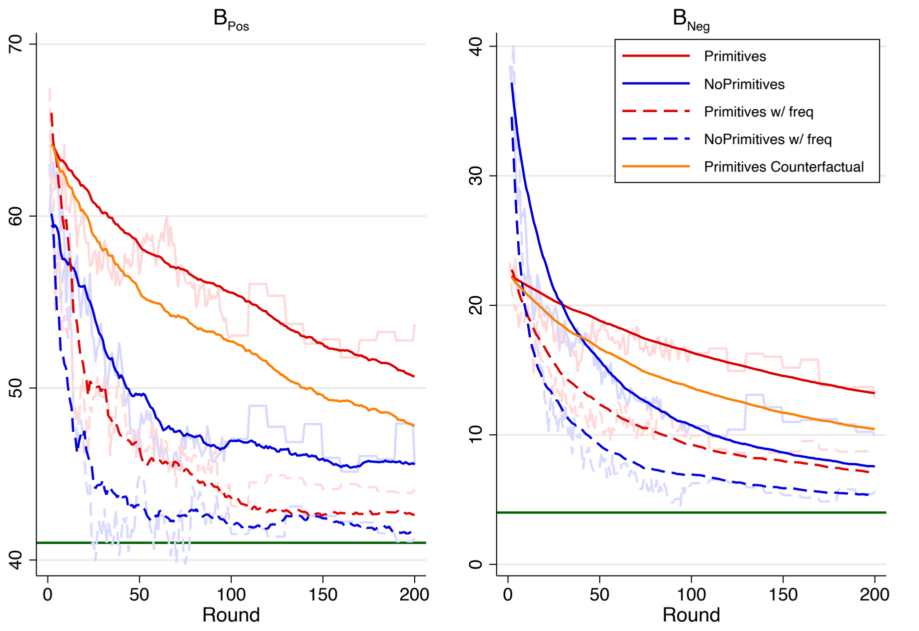


Figure 10: Estimates of the Learning Model for Treatments with Frequencies and Core Treatments

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.

hypothesis that confidence in an initially incorrect answer can hinder learning from feedback by making people less responsive to new information.

Nonetheless, our estimates for $\sigma$ 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 where feedback is presented in an aggregated, processed form, our estimates for $\sigma$ are twice as high (for both $B_{Pos}$ and $B_{Neg}$) in *NoPrimitives* relative to *Primitives*.[45]

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

[45] Estimates of $\sigma$ for $B_{Pos}$ are 0.10 and 0.18 in *Primitives* and *NoPrimitives*, respectively. The corresponding values are 0.19 and 0.35 for $B_{Neg}$. Statistical tests using bootstrapping show differences to be significant (p-value $< 0.000$ in both cases).

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 10. 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. This analysis reveals that, by round 200, approximately half the difference between *Primitives* and *NoPrimitives* is due to lower attentiveness in the former treatment.

***Result #4:*** *Beliefs move closer to the Bayesian benchmark when feedback is presented in a processed way. Minimizing cost of attentiveness in this way eliminates the treatmentment difference between environments where primitives are provided vs. not provided (as seen between Primitives w/ freq and NoPrimitives w/ freq). Counterfactual analysis using a learning model suggests approximately half the difference observed by round 200 between our core treatments (Primitives vs. NoPrimitives) to be driven by differences in attentiveness to feedback, with the rest coming from differences in confidence.*

### 4.4 Transfer learning: Behavior with different primitives

Our results in the previous sections establish that suboptimal behavior can be corrected in the long run in treatments with primitives when feedback is presented in a processed, easy-to-read format. However, so far our design does not distinguish between two different ways in which subjects in such settings 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). In turn, this type of deeper learning can help improve decision making in related but different environments (where subjects cannot rely on previously accumulated data). Our goal in this section is to provide further insights on exactly what subjects learn from their experiences and whether learning can be transferred when primitives change.[46]

We tackle this question in the last (fourth) part of our core treatments, where subjects face a new updating task in which the primitives are changed to $p' = .95$ and $q' = .85$. Prior to this part,

[46] 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 part three, we presented subjects in these treatments with ample feedback processed for them such that almost all subjects converged to beliefs very close to the Bayesian benchmark.[47]

In this last part of the experiment, subjects in the core treatments are asked to report beliefs just once, without any feedback. We call this final round with new values $p'$ and $q'$ "round$_{(p',q')}$" and assess how beliefs differ from the Bayesian benchmark relative to answers in round one and an appropriate control, as described below.

On average, beliefs in *Primitives* for round$_{(p',q')}$ are $(B_{Pos}, B_{Neg}) = (85, 41)$, where the benchmarks are $(B^{Bay'}_{Pos}, B^{Bay'}_{Neg}) = (99, 77)$ for Bayesian updating and $(B^{pBRN'}_{Pos}, B^{pBRN'}_{Neg}) = (85, 15)$ for perfect base-rate neglect, pBRN. But, as the distribution of responses plotted in Figure 11 reveals, there are essentially two clusters of responses; one cluster is close to pBRN and the other to the Bayesian benchmark. A comparison with the left plot of Figure 4 suggests that subjects are providing beliefs closer to the Bayesian posterior in round$_{(p',q')}$ relative to round one. Specifically, 17.2 percent of subjects provide pBRN beliefs in this part, a share that is below the 56.2 percent observed in round one. Meanwhile, 12.5 percent of subjects provide beliefs exactly at the Bayesian benchmark compared to 4.7 percent in round one.

*NoPrimitives* acts as a control allowing us to interpret behavior in round$_{(p',q')}$ in *Primitives.* In particular, subjects in *NoPrimitives* were also told the primitives $p' = .95$ and $q' = .85$ for round$_{(p',q')}$. Thus, subjects in both treatments faced this final round in an identical manner. However, subjects in *NoPrimitives*, unlike most subjects in *Primitives* did not go through an experience where they used information on the primitives to form an initial answer, and then through feedback they accumulated realized the suboptimality of this answer. Hence, subjects in *Primitives* uniquely had the opportunity to correct how they make use of information on primitives, possibly learning to overcome the dominant mistake on BRN.

Average beliefs in *NoPrimitives* for round$_{(p',q')}$ are $(B_{Pos}, B_{Neg}) = (30, 81)$. That is, while the average for $B_{Pos}$ is similar across treatments, there is a significant 11 percentage point difference in terms of $B_{Neg}$ (p-value 0.028), which is the dimension on which the Bayesian and pBRN benchmarks differ the most, with the belief being closer to the Bayesian benchmark in *Primitives.* However,

[47] Specifically, subjects were given 1000 rounds of feedback aggregated in two-by-two table with empirical frequencies of success conditional on a positive and negative signal computed. See Online Appendix B for more details on implementation. Density plots in Online Appendix C.2 show convergence to the Bayesian benchmark after these interventions. 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.

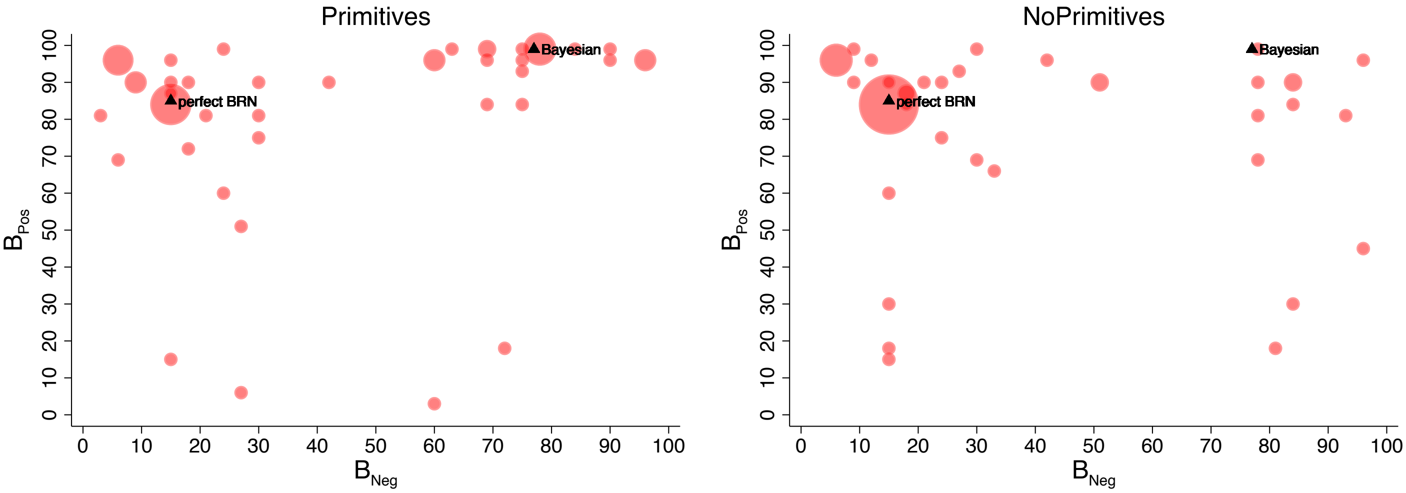


Figure 11: Transfer Learning: Density Plots in Final Round with New Primitives

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. The data is from the final round of the core treatments where the prior and the reliability of the signal is changed.

the difference is more clearly visible in the distribution of beliefs across treatments as presented in Figure 11.

In *NoPrimitives* there are almost no subjects around the Bayesian benchmark (1.6 percent), but a relatively larger group is concentrated around the pBRN point (37.5 percent). In fact, 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. Meanwhile, a similar exercise does not change the proportion of subjects submitting Bayesian beliefs in *NoPrimitives* (still 1.6), but it increases to 15.6 in *Primitives*. Average beliefs in *Primitives* are significantly closer to the Bayesian benchmark relative to *No Primitives* (p-value 0.022).

This treatment effect (which switches direction relative to earlier parts!) suggests 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 continue to be far from the Bayesian benchmark.

***Result #5:*** *When subjects are exposed to different but known primitives, the treatment effect (between Primitives and NoPrimitives) reverses: Subjects in NoPrimitives now neglect the prior to a significantly larger extent than subjects in Primitives. This suggests that some subjects in*

*Primitives can learn to take the prior into account when facing new primitives.*

## 5 Discussion and 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 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.[48]

[48] Details about experimental design are presented in Online Appendix B.

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.[49] Subjects who fail to do so might incorrectly perceive this as a choice reflecting their risk preference, endowing them with confidence 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.[50]

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).[51] 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

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

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

[51] 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.

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)*.

These treatments were conducted on Prolific with 130 subjects per treatment and 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 $< 0.000$ 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 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)*.[52] 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.[53]

[52] 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 G.

[53] 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

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.

We also recruited 130 subjects on Prolific for these treatments and results 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 depending on whether subjects make an optimal round-one choice or not.[54] 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:*** *By the last round, behavior is more optimal in Primitives (Voting) relative to NoPrimitives (Voting). 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 as reflected in the comparison between Complex Primitives (Voting) and Complex NoPrimitives (Voting).*

## Discussion

These results show in a new setting how suboptimal behavior can be persistent in the long run even in information-rich environments. Note that learning from feedback for an attentive subject is trivial in our implementation of the voting problem: Of the two actions available to the subject, feedback reveals outcomes of $6 for one, but $6 or $10 for the other. Reinforcing our earlier results from the updating problem, we find that initial misconceptions, which drive suboptimal behavior, can also prevent learning from feedback by impacting an agent's responsiveness and attentiveness to this information. Results from the complex framing of the voting problem—with higher rates of optimal

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).*

[54] 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).* This difference is likely underestimating the real difference because the evidence suggests that learning in the Complex case is more challenging; see Online Appendix G.

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 |
| $\Delta$ | 17.7 | -5.3 | 27.3 | -1.21 | -1.5 |
| p-value | 0.003 | 0.495 | 0.001 | <0.000 | <0.000 |
| *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 |
| $\Delta$ | 3.1 | -8.4 | 11.9 | -0.63 | 0.00 |
| p-value | 0.584 | 0.248 | 0.128 | < 0.000 | 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.

behavior and no treatment difference with respect to whether primitives are provided—reaffirm the idea that confidence in initial answer plays a key role in hindering learning from feedback.

These results, combined with earlier findings from the updating task, 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 satisfycing (Caplin, Dean & Martin 2011), but harder in settings where suboptimal behavior is driven by conceptual mistakes subjects 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.[55] Confidence measures in initial responses can be useful 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.[56]

From a policy perspective, an important implication of our results is that for interventions designed to counter systematic biases to succeed, they need to move beyond providing information that is indicative of optimal behavior and target agents' engagement with this information. The results also reveal several counterintuitive interventions that can be 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.

Finally, we see several directions in which this research agenda can be advanced. First, our paper focuses primarily on failures in a simple updating problem, with base-rate neglect being the dominant mistake. However, as demonstrated with our supporting treatments using a voting problem, the experimental design we propose can be incorporated to other settings to study the persistence of other well-documented biases in response to different forms of feedback. Adopting this approach more broadly can help better identify what types of biases are persistent even in information rich environments. Second, there are other channels through which initial misconceptions can prevent learning from feedback. We focus on simple decision problems where feedback was exogenous to an agent's decision. Moving beyond this paradigm—looking at games and decision problems with endogenous feedback—would uncover other forces that contribute to the persistence of suboptimal behavior. Finally, while the controlled environment the laboratory provides is a natural starting

[55] 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-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.* We also find that in the initial rounds of a treatment that can trigger incorrect heuristics (*Primitives*) very few subjects make optimal choices. Our focus, however, is on the extent to which informative feedback can correct suboptimal choices in the long run. If in the BSR elicitation problem subjects are not aware that they are making suboptimal choices, our results suggest that informative feedback would not help them very much in the long run.

[56] See Enke & Graeber (2022) for a cognitive uncertainty measure, and Enke, Graeber & Oprea (2022) for evidence on how confidence varies among some of the most well known biases in behavioral economics.

point to study the interaction between biases and learning, we believe that it is important to assess the extent to which biases persist in prominent field applications. For example, future work can study base-rate neglect in doctors interpreting medical tests using types of feedback that are natural in that setting.

# 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.

Arkes, H. R. & Blumer, C. (1985), 'The psychology of sunk cost', *Organizational behavior and human decision processes* **35**(1), 124–140.

Bar-Hillel, M. (1980), 'The base-rate fallacy in probability judgments', *Acta Psychologica* **44**(3), 211–233.

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

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., Bodoh-Creed, A. & Rabin, M. (2019), 'Base-rate neglect: Foundations and implications', *Working Paper* .

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.

Cox, J. C. & Oaxaca, R. L. (2000), 'Good news and bad news: Search from unknown wage offer distributions', *Experimental Economics* **2**(3), 197–225.

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. (2022), Cognitive uncertainty, Technical report, National Bureau of Economic Research.

Enke, B., Graeber, T. & Oprea, R. (2022), Confidence, self-selection and bias in the aggregate, Technical report, National Bureau of Economic Research.

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. (2021), 'Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory', *Working paper* .

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. & 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.

Gigerenzer, G. & Hoffrage, U. (1995), ‘How to improve bayesian reasoning without instruction: frequency formats.’, *Psychological review* **102**(4), 684.

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.

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.

James, G. & Koehler, D. J. (2011), 'Banking on a bad bet: Probability matching in risky choice is linked to expectation generation', *Psychological Science* **22**(6), 707–711.

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).

Kahneman, D. & Tversky, A. (1973), 'On the psychology of prediction.', *Psychological review* **80**(4), 237.

Koehler, D. J. & James, G. (2009), 'Probability matching in choice under uncertainty: Intuition versus deliberation', *Cognition* **113**(1), 123–127.

Koehler, D. J. & James, G. (2010), 'Probability matching and strategy availability', *Memory & cognition* **38**(6), 667–676.

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.

Mobius, M. M., Niederle, M., Niehaus, P. & Rosenblat, T. S. (2022), 'Managing self-confidence: Theory and experimental evidence', *Working paper* .

Moore, D. A. & Healy, P. J. (2008), ‘The trouble with overconfidence.’, *Psychological review* **115**(2), 502.

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.

Rabin, M. (2000), ‘Inference by believers in the law of small numbers’, *Quarterly journal of Economics* **117**(3), 775–816.

Schotter, A. & Braunstein, Y. M. (1981), ‘Economic search: an experimental study’, *Economic inquiry* **19**(1), 1–25.

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.

Thaler, R. (1980), ‘Toward a positive theory of consumer choice’, *Journal of Economic Behavior & Organization* **1**(1), 39–60.

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.