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

Ignacio Esponda
UCSB

Emanuel Vespa
UCSD

Sevgi Yuksel
UCSB

May 25, 2023

Abstract

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

*We would like to thank the editors, anonymous referees and numerous seminar participants for helpful comments. We also wish to thank Dingyue Liu and Caroline Zhang for excellent research assistance. 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. An important question is whether such biases are self-corrected in the presence of feedback. On the one hand, biases might vanish with experience if agents are accumulating evidence informative of optimal behavior. On the other hand, this type of learning presumes agents are attentive to the feedback they are experiencing, willing and able to adjust their behavior in response to it. A growing empirical and theoretical literature challenges this position by emphasizing how initial misconceptions can have long-lasting effects on how people learn from their experiences.[1]

In this paper, we present results from a laboratory experiment designed to study optimality of long-run behavior in the presence of feedback and bring to light the different mechanisms that hinder learning from feedback. The experiment has two crucial features. First, we consider a baseline treatment where subjects face a decision problem in which useful information about the problem generates biased behavior. We then study the evolution of this bias when subjects face multiple rounds and receive transparent feedback. Second, we compare behavior in this treatment to a control treatment in which information inducing biased behavior is withheld from the subjects. In the absence of such information, subjects can only 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.

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

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

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

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

In each round, a new state is randomly selected and a signal is drawn. Subjects submit beliefs conditional on the signal, and observe the true state at the end of the round. The interface also displays a record of all past outcomes. In our baseline treatment, labeled as *Primitives*, subjects are presented with the above problem (albeit with a more neutral framing) and informed of the primitives (i.e., the 15 percent prior and the 80 percent accuracy of the signal) so that, in principle, they could provide the correct response of 41 percent (conditional on a positive signal) from the very first round.

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

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

We find an important treatment effect after 200 rounds with respect to the accuracy of beliefs: In aggregate, beliefs in the control treatment (*NoPrimitives*) are closer to the Bayesian benchmark relative to beliefs in the baseline treatment (*Primitives*). For example, the average belief conditional on a positive signal is at 46 percent in *NoPrimitives* which is eight percentage points lower than the

value in *Primitives*.[3] Moreover, the treatment effect disappears if we exclude subjects who provide the pBRN answer in the initial round, suggesting that, of all initial misconceptions induced by the *Primitives* treatment, it is principally those inducing the pBRN beliefs in round one that hinder learning from feedback.

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

First, we investigate whether initial misconceptions, induced by information on the primitives, hinder learning from feedback by endowing subjects in *Primitives* with unjustified high confidence in their initial responses. To test this, we run a diagnostic treatment that is identical to *Primitives* except for one small difference. At the end of round one, we tell subjects (to whom the message applies) that their initial responses are not correct. Otherwise, subjects experience 200 rounds of feedback in the same way. The message has a large impact on how close beliefs are to the Bayesian benchmark after 200 rounds of feedback. The average belief conditional on a positive signal drops to 42 percent, 12 percentage points lower than the value in *Primitives*. In fact, all subjects, including those with initial pBRN beliefs, are capable and willing to learn from feedback. Together with our earlier findings, this suggests that subjects with high confidence in their initial pBRN beliefs play a critical role in inhibiting learning from feedback in *Primitives*.

Second, we ask whether initial misconceptions, induced by information on the primitives, hinder learning from feedback by also reducing subjects' attentiveness to the feedback available to them.[4] Specifically, we conduct a set of diagnostic treatments to examine whether information on the primitives impacts engagement with feedback. These treatments are identical to *Primitives* and *NoPrimitives*, except that we allow subjects to "lock in" their responses at any point during the 200 rounds. Once responses are locked-in, they are automatically implemented for all future rounds. This lock-in decision gives us a simple measure of engagement by revealing how many rounds of feedback subjects are willing to see. Our results highlight large differences in engagement with feedback. When provided with the primitives, only half the subjects choose to see more than 20 rounds of feedback and only four percent choose to observe all 200 rounds. By contrast, without

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

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

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.

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

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

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

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

attentiveness to such information.

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

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

Connections to the literature

The themes explored in this paper, in terms of how learning from past experiences is necessarily shaped by our initial understanding of the world, connect with a few different literatures. First, our results provide support for a growing literature in economics that studies the implications of incorrect or misspecified models. A central premise of this literature is that the degree to which an

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

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

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

Even in the absence of direct information-processing costs, there could be other behavioral forces that influence an agent's engagement with feedback. For example, either due to motivated beliefs (e.g. Bénabou & Tirole 2003; Brunnermeier & Parker 2005; Köszegi 2006) or simply due to a desire for consistency (Falk & Zimmermann 2018), agents might be reluctant to adjust their

[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 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.

behavior in response to past outcomes.[10] These different literatures share a common insight that initial misconceptions can inhibit learning by impacting the way agents engage with the data, and our experiment provides strong evidence for this channel.

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

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

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

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

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

[13] There is also a literature related to the voting problem that we study in our last treatments. As a reference, see Esponda and Vespa (2014, 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.

2 Experimental design

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

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

I. Updating task: Round One

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

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

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

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

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

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

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

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

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

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

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

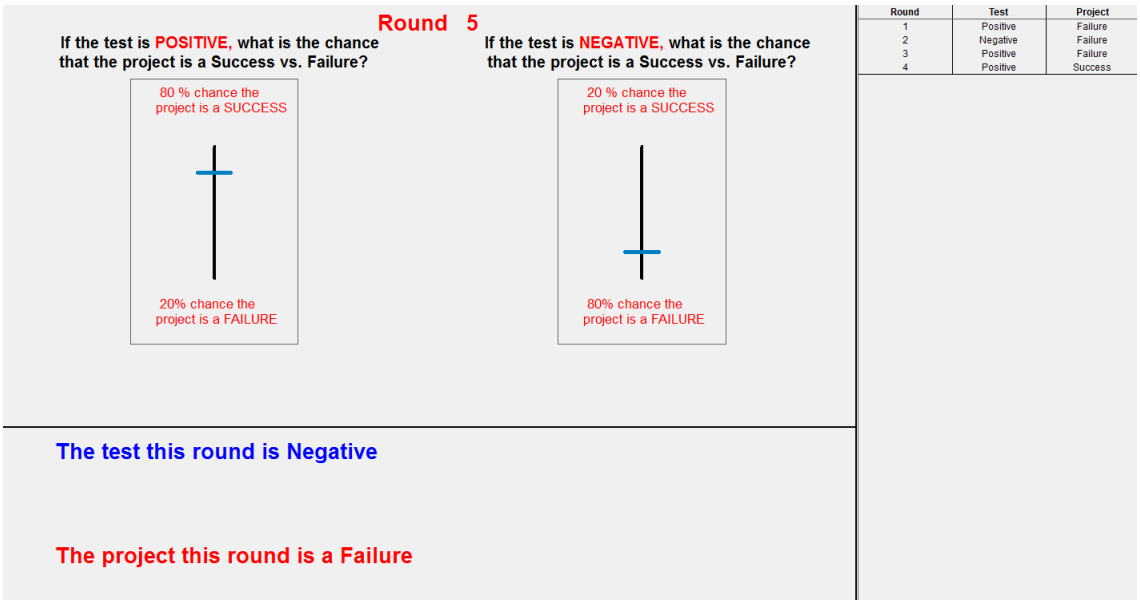


Figure 1: Interface Screenshot at Round Five of Core Treatments

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

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

Experimental procedures

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

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

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

3 Results on *Primitives* vs. *NoPrimitives*

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

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

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

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

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

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

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

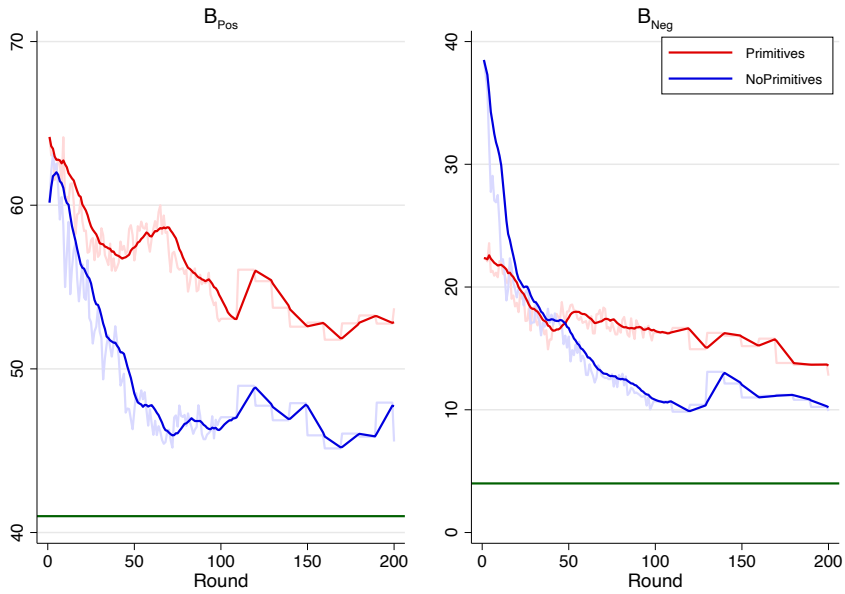


Figure 2: Evolution of Beliefs in *Primitives* and *NoPrimitives*

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

3.2 Learning in *Primitives* vs. *NoPrimitives*

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

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

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

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

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

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

3.3 Heterogeneity

To provide an overview of the heterogeneity in responses, Figures 3 and 4 present the distribution of beliefs in *Primitives* and *NoPrimitives* at the initial and final rounds. As mentioned earlier, most subjects (56.3 percent) submit beliefs consistent with pBRN in round one of *Primitives*. By round 200, however, the distribution of beliefs in *Primitives* has shifted significantly, with one large cluster close to or at the pBRN point and another one close to or at the Bayesian point. In fact, 12 percent of subjects submit beliefs consistent with pBRN in both round one *and* round 200.[26]

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

These patterns suggest long-run differences between *Primitives* and *NoPrimitives* to be possibly driven by those subjects who initially display perfect base-rate neglect in *Primitives*. To begin to

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

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

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

[26]By round 200, 35 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.

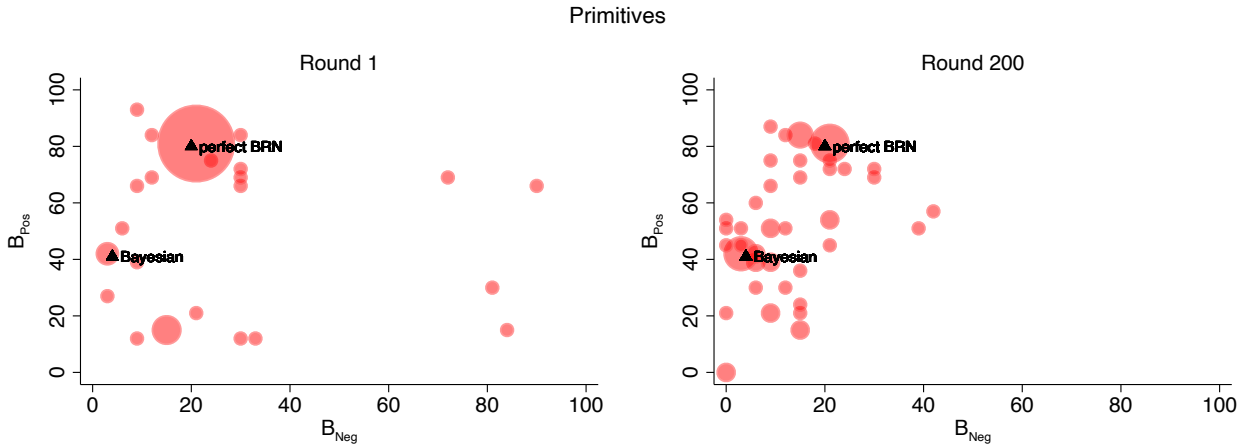


Figure 3: Density plots for *Primitives*

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

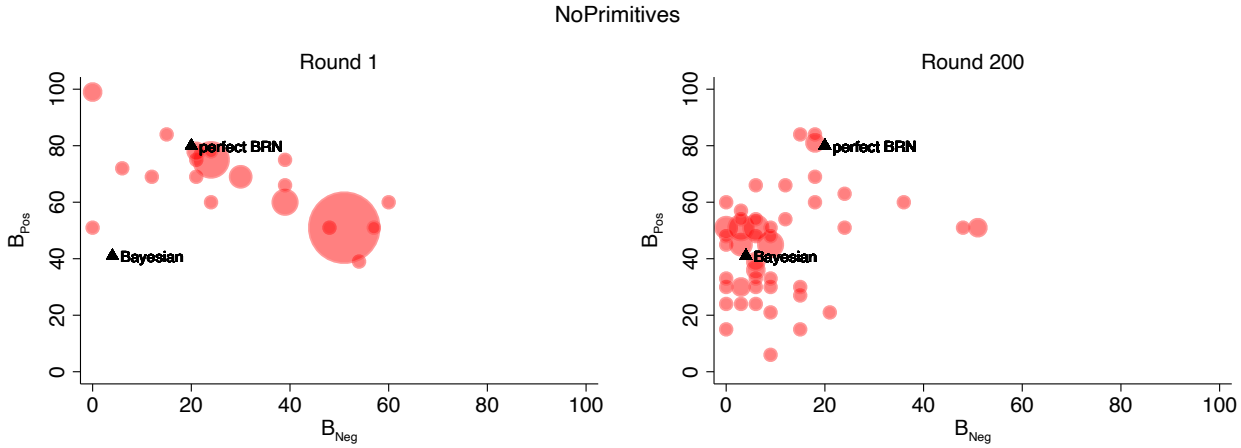


Figure 4: Density plots for *NoPrimitives*

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

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

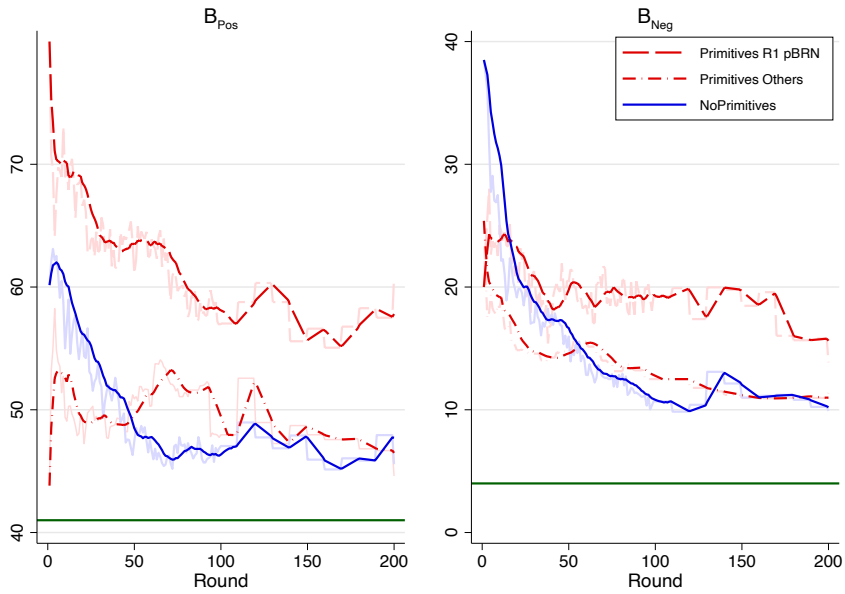


Figure 5: Evolution of Beliefs for R1 pBRN Subjects and Others in *Primitives* vs. *NoPrimitives*

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

all others (i.e., non-pBRN subjects), however, are not different from the average beliefs of subjects in *NoPrimitives* (p -value 0.760).[27]

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

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

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

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

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

4 Mechanisms

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

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

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

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

4.1 Confidence

If confidence in an incorrect initial answer is the reason why subjects don't learn as effectively in *Primitives*, then a shock to their confidence should facilitate learning. To test this possibility, we conduct a new treatment, *Primitives w/ shock*, that is identical to *Primitives* except for one

difference: If a subject submits an incorrect answer in round one, the computer interface sends them a message that says that their answer is incorrect before they start with round two.[30]

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

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

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

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

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

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

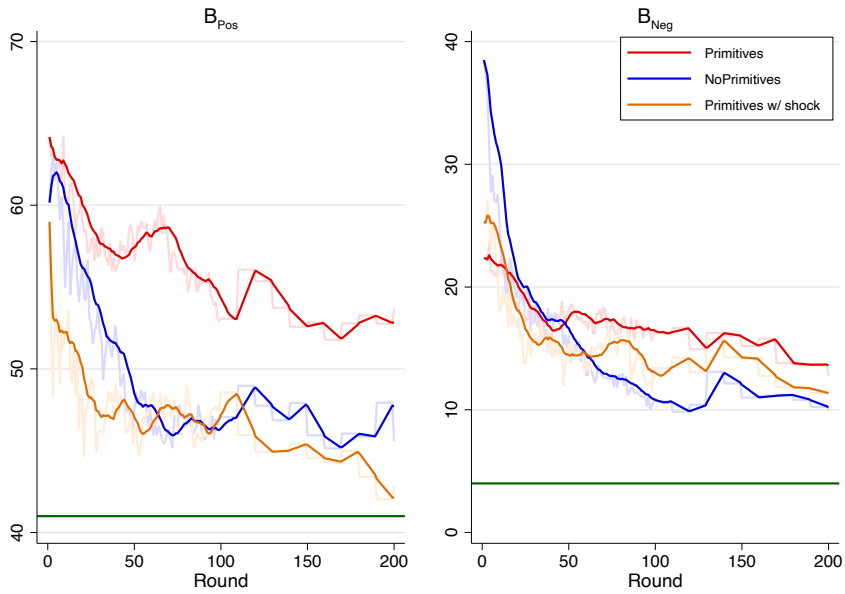


Figure 6: Comparing Evolution of Beliefs in *Primitives w/ shock* to *Primitives* and *NoPrimitives*

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

4.2 Attentiveness

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

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

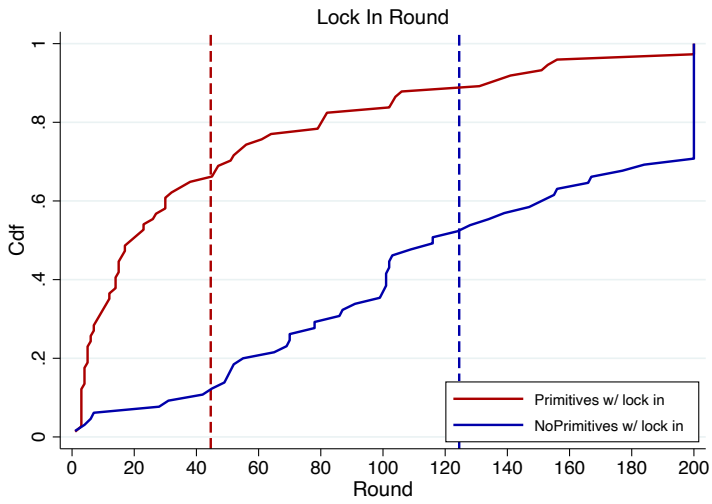


Figure 7: Distribution of Lock In Decisions

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

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

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

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

[34]In Online Appendix E, we confirm that initial responses are similar between the core treatments and the new ones with the lock in option. One difference is that there are slightly fewer pBRN subjects in *Primitives w/ lock in* relative to *Primitives*: 42 percent vs. 56 percent (p -value 0.094). 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.

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

Primitives w/ lock in, only half the subjects choose to see more than 20 rounds of feedback and only four percent of subjects choose to see all rounds of feedback. By contrast, in *NoPrimitives w/ lock in*, 94 percent of subjects choose to see more than 20 rounds of feedback and 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 -value < 0.001).

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

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

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

4.3 Costly attention

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

steps. We first test the null hypothesis that the distribution in *NoPrimitives w/ lock in* either first order stochastically dominates or is equal to the distribution in *Primitives w/ lock in*. We reject this null hypothesis (p-value <0.001). We then test the null hypothesis that the distribution in *Primitives w/ lock in* first order stochastically dominates the distribution in *NoPrimitives w/ lock in*. We cannot reject the null in this case (p-value 0.829).

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

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

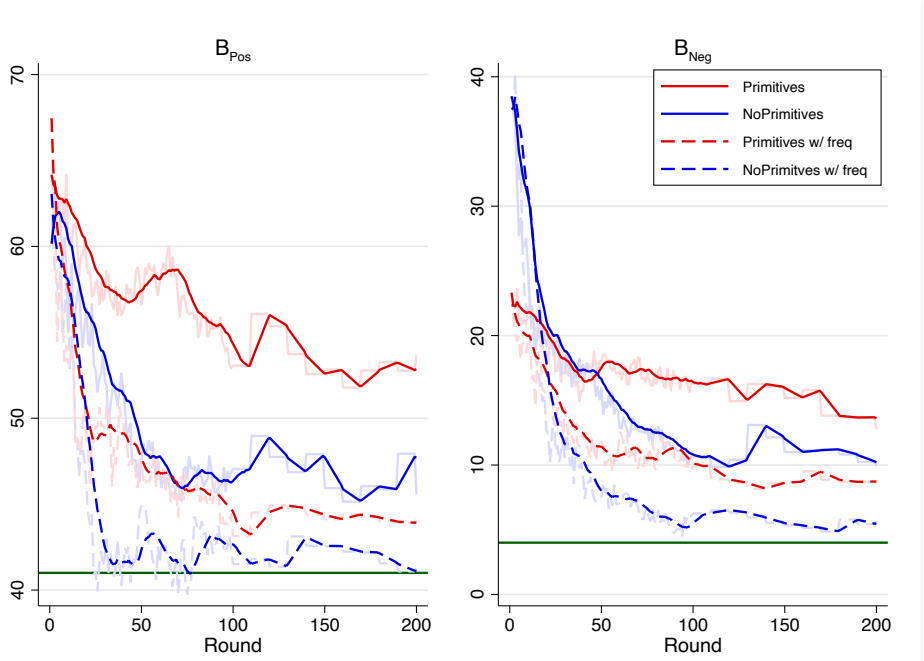


Figure 8: Evolution of Beliefs in Treatments with Frequencies Relative to Core Treatments

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

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

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

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

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

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

4.4 A model of learning

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

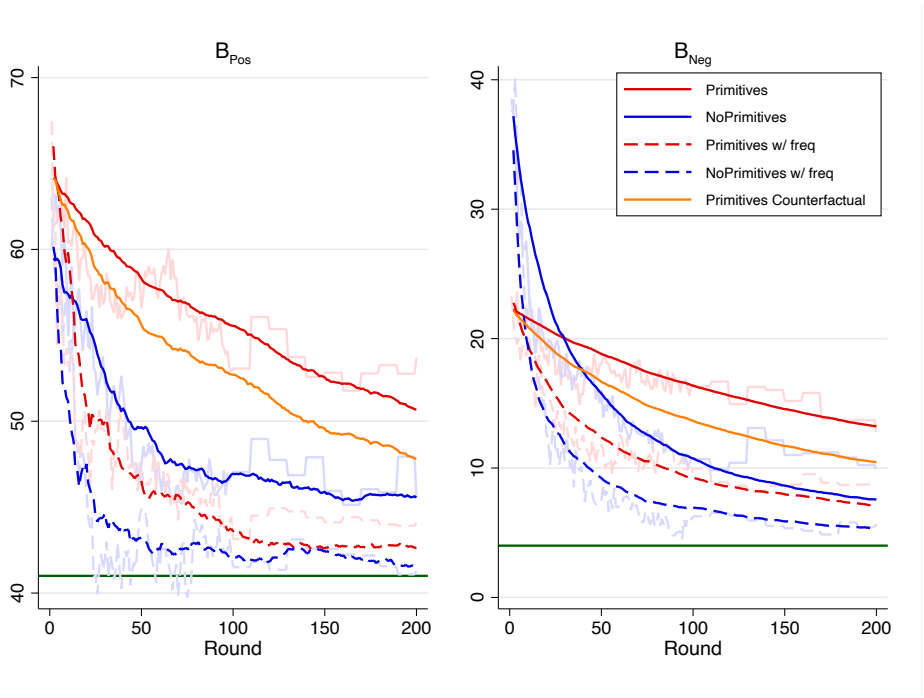


Figure 9: Estimates of the Learning Model for Treatments with Frequencies and Core Treatments

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

to estimate σ.[44]

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

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

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

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

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

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

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

4.5 Transfer learning: Behavior with different primitives

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

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

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

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

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

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

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

5 Evidence beyond the updating problem

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

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

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

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

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

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

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

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

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

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

in their suboptimal choice.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

	Optimality of Vote in Last Round (in %)			Confidence	
	All	R1 Optimal	R1 Not Optimal	Mean	Median
Primitivites (Voting)	56.9	84.1	43.0	3.76	4.00
NoPrimitivites (Voting)	74.6	78.8	70.3	2.55	2.50
Δ	17.7	-5.3	27.3	-1.21	-1.5
p-value	0.003	0.495	0.001	<0.001	<0.001
Complex Primitivites (Voting)	70.0	87.2	57.3	3.39	3.00
Complex NoPrimitives (Voting)	73.1	78.8	69.2	2.76	3.00
Δ	3.1	-8.4	11.9	-0.63	0.00
p-value	0.584	0.248	0.128	< 0.001	1.00

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

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

6 Conclusion

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

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

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

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

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

References

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

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

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

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

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

[58] See Enke & Graeber (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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Enke, B. & Graeber, T. (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* .

Esponda, I., Vespa, E. & Yuksel, S. (forthcoming), Mental models and transfer learning.

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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).

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.

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

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

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

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

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

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

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

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

Online Appendix for

Mental Models and Learning:

The case of Base-Rate Neglect

Ignacio Esponda Emanuel Vespa Sevgi Yuksel

CONTENTS:

A Literature review: Further details

A.1 Literature on base-rate neglect with feedback

This section first provides a review of the experiments on base-rate neglect (BRN). Our focus is on the extent to which the different studies document changes in behavior in response to feedback. At the end of the section we also include a brief overview of probability-matching experiments and the connection to our paper.

The literature on base-rate neglect is founded on two seminar papers by Kahneman and Tversky (1972, 1973). The two papers differ in the type of updating problem used in the experiment to study base-rate neglect. In Kahneman and Tversky (1973) subjects were asked to make a judgment about the probability that a person is an engineer or a lawyer based on a description. The description provided was designed to include characteristics "representative" of being either an engineer or a lawyer.[59] However, this design was criticized by some (Nisbett et al. 1976) who were concerned that the detailed textual description provided as a signal, which stood in contrast to the statistical description of the prior, could explain why base rates were not as strongly incorporated into posterior beliefs. However, base-rate neglect is also observed in more standard updating problems. Kahneman & Tversky (1972) purposefully used an abstract problem (although framed as the famous cab problem), where the state and signal were simply colors (green vs. blue) and the reliability of the signal was explicitly given to the subjects to enable Bayesian updating.[60] The parameters used in our experiment are precisely the values from this paper, although we change the framing slightly as described in the experimental-design section. The literature that followed from these papers broadly falls into two corresponding categories: experiments where the primitives are fully provided (as in Kahneman & Tversky 1972) or experiments where either the prior or the signal reliability is open to interpretation (as in Kahneman and Tversky 1973).

[59] After being provided with a prior (on the person being a lawyer of an engineer), subjects were given, for example, the following description. "Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles." Results revealed subjects' posteriors to vary very little with the base rate. An important advantage of this design is that the degree to which base rates are incorporated into the posterior can be tested without explicitly fixing the informativeness of the description (hence, without studying directly whether subject over or under react to the information).

[60] Subjects were asked the following problem: "Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and the remaining 15 percent are Green. A cab was involved in a hit-and-run accident at night. A witness later identified the cab as a Green cab. The court tested the witness' ability to distinguish between Blue and Green cabs under nighttime visibility conditions. It found that the witness was able to identify each color correctly about 80 percent of the time, but confused it with the other color about 20 percent of the time. What do you think are the chances that the errant cab was indeed Green, as the witness claimed?" The correct answer is 41percent.

Grether (1980, 1992) and Griffin and Tversky (1992) are some of the early economics-style experiments on the topic where subjects are financially incentivized to form accurate beliefs and the updating problems are presented in the standard framework of judging the likelihood of abstract events (for example, event involving balls drawn from different urns). Importantly, Grether (1980) also introduces a general way of measuring partial base-rate neglect based on regression analysis focusing on the log likelihood ratio of different events. This approach is now commonly used in many papers, including this one, studying updating behavior. It should be noted that none of these early papers studied how behavior changes with feedback. In most experiments subjects only answered one belief updating question, and in others that included multiple questions, the parameters and/or the environment changed between questions with no feedback between questions.

The literature on base-rate neglect grew quickly in the next few decades. Koehler (1996) provides an extensive review of experiments on base-rate neglect up to that point. There are three important observations in this paper that are relevant to our research question. First, Section 2.1.1 of this paper concludes that in experiments where subjects are faced with multiple versions of a belief elicitation question (without any feedback) whether the base rate or the characteristics of the signal are varied within subject can have an impact of the results. In general, subjects respond more to base rates if they are varied within, or alternatively if there is no variation in signal characteristics within. Second, the paper highlights a line of research studying whether the base rate is integrated more in a belief updating problem when the question is framed or presented in terms of frequencies rather than probabilities. This perspective was first introduced by Gigerenzer (1991) and Gigerenzer & Hoffrage (1995). Further evidence on different aspects of this are also presented in Cosmides & Tooby (1996), and more recently in Barbey and Sloman (2007).

Third, more closely related to our research question, Section 2.1.2 of Koehler (1996) discusses several early experiments where subjects have an opportunity to learn about base rates from direct feedback. For example, Manis et al. (1980), Lindeman et al. (1988), and Medin and Edelson (1988) provide evidence that base rates influence probabilistic judgements more when they are directly experienced through trial-by-trial outcome feedback. None of these papers include a treatment that can be mapped back cleanly to either of our treatments, but they provide insights that parallel some of our findings. In Manis et al. (1980) subjects were shown 50 yearbook pictures of male students and, for each randomly selected picture, they were asked to predict the person's position on two issues (marijuana legalization and mandatory seatbelt legislation). Note that a signal in this context can be interpreted to be the characteristics of person observed in the picture. The informativeness of these pictures is ambiguous and actually manipulated to be non-existent. The results suggest that

subjects adjust their judgments in response to the accuracy of their past predictions. In Lindeman et al. (1988) subjects are given 16 different versions of Kahneman and Tversky's engineer-lawyer problem. While the analysis indicates that feedback leads to adjusted probability estimates closer to the Bayesian benchmark, the type of feedback that subjects are provided is highly unnatural and unusual.[61] It is also important to note that the paper does not find any transfer of learning in this environment to another one where subjects can display base-rate neglect (based on Zukier and Pepitone 1984). Medin and Edelson (1988) report results from an experiment where the task involved participants diagnosing hypothetical diseases on the basis of symptom information. It is difficult to interpret their results as their learning environment is complicated by the fact that there are many features of the environment that are varied within subjects and some of these involve ambiguous signals. Overall, they find mixed results for subjects incorporating the base rate. Among these set of papers, the closes to our work is Christensen-Szalanski and Beach (1982). The paper demonstrates that subjects make use of base rates in forming posterior probabilities when they have experienced the relationship between the base rate and the diagnostic information, but fail to make use of the base rate when they only experience the base rate and are given the reliability of the signal.[62]

Since the review article of Koehler (1996), there has been a considerable literature in psychology studying whether subjects can learn through direct experience to incorporate base rates into posterior beliefs. These papers are reviewed in Goodie and Fantino (1999). While this body of work often provides evidence that subjects can learn from experience to adjust actions towards optimal behavior, the approach in these papers are fundamentally different from ours. The framework adopted in most of these experiments is one where subjects repeatedly choose between two binary options after observing a binary cue, receiving feedback about the optimality of the choice after each round. The choices are often between abstract options (for example, green or blue) and the cues could be labeled similarly or differently from the options (for example, matching colors or arbitrary shapes). Critically, subjects are not informed about the primitives determining statistical

[61] In each problem, subjects were asked to form beliefs based on the same description using different base rates. While the informativeness of the description is not explicitly given in this experiment, a subject's answer to the first question implies a 'correct' answer to the second question if subjects are assumed to be Bayesian. The experiment elicited both beliefs while giving feedback on what the 'correct' answer should have been to the second question (conditional on the answer to the first question).

[62] One of their treatments (where subjects experience both the state and the signal in direct feedback) is similar to our *NoPrimitives* treatment where subject are not given the primitives and learn from feedback. However, a critical difference is that subjects form beliefs only *after* observing *all* the feedback. Christensen-Szalanski and Beach (1982) also go further and tell subjects explicitly that they "will be asked to use this information" to answer several question in the future. In their second treatment, they provide subjects only with the reliability of the signal, and then provide subjects with 100 rounds of natural feedback only on the base rate. They find that subjects cannot successfully make use of the feedback in this context.

relationship between the cue and the optimal action.[63] In this respect, these experiments are closest to our *NoPrimitives* treatment in which the prior and the reliability of the signal were not provided to the subjects. However, there are still some differences in how such a treatment is implemented in these papers that could be important for behavior. For example, in these experiments, subjects are not told explicitly that the environment they face repeatedly is a stationary one in the sense that each round corresponds to an independent draw of optimal action/cue pair from the same distribution. Note also that the learning problem is different from the one we study in that in these experiments subjects can possibly learn the optimal binary action conditional on each signal without ever forming precise beliefs conditional on each signal.

Despite the relatively large literature on the topic, we have not identified a paper that includes a treatment in which subjects were provided with the primitives and also had to opportunity to learn from direct feedback while repeatedly experiencing the same environment. Moreover, we have not found a single study that compares differences between the description and experience paradigms within the same sample of subjects.[64] Fantino and Navarro (2012) provide a survey of the description-experience gap (the finding that people respond differently to the same quantitative information depending on whether it is described or experienced) in different environments. With respect to the description-experience gap in base-rate neglect experiments, they compare across experiments within each paradigm (only description experiments, such as Kahneman & Tversky (1972), or only experience experiments, such as Goodie and Fantino (1996)). That is, they report that there was no single study that compared the description to the experience paradigm within the same group of participants.

Literature on probability matching & feedback

The experimental literature on probability matching is surveyed in Vulkan (2000) and, more recently in Erev and Haruvy (2013). Most papers in the early literature on probability matching used an environment in which the primitives were not provided to subjects. To illustrate, here is a typical example taken from Erev and Haruvy (2013). There is an event E that happens with probability 0.7, but subjects do not know this probability. In a given round, subjects click on button H or

[63] In these experiments subjects are not even allowed keep track of past realizations. In the instruction subjects are explicitly told: "Please don't use any outside tools, such as a pencil and paper, to help you remember what you saw" (Goodie and Fantino 1999).

[64] The 'experience' paradigm corresponds to experiments described in the previous paragraph (surveyed in Goodie and Fantino (1999)), where subjects are not provided with the primitives but can learn from feedback. Meanwhile, the 'description' paradigm captures the standard Kahneman & Tversky (1972) example, where primitives are provided and subjects answer one question. Notice that this comparison does not involve a treatment in which people are given the primitives *and* feedback.

button L. Button H pays (L pays) a positive amount if E occurs (E does not occur). After 50 rounds, the observed rate of H selection is 70%. This finding coincides with the earlier literature in which subjects were reported to make choices that are close to 'probability matching' instead of optimizing. However, more recent papers have demonstrated that longer experience slowly moves choices toward maximization. In that example, the H rate in the data was 90% between rounds 51 and 150. These findings are consistent with our long-run findings for the *NoPrimitives* treatment.

Relatively recent papers do provide primitives.[65] Newell et al. (2013) present an experiment in which a 10-sided die with 7 green and 3 red sides, which subjects can see, is going to be rolled in each round. The subjects' task is to predict the color of the die. In the first 50 rounds, the rate of green choices was close to 80 percent. In rounds 51 to 150, the rate is close to 85 percent. These findings suggest that while subjects make choices consistent with probability matching early on, suboptimal choices decrease with feedback. This finding is in line with our result for the *Primitives* treatment in which feedback moves average beliefs closer to the Bayesian benchmark.

Koehler & James (2010) provide evidence suggesting that when primitives are provided, the 'probability-matching' heuristic more readily comes to mind relative to the optimal strategy. This opens up the possibility that subjects may have confidence in an incorrect choice, but we did not find a reference that would measure confidence. While the evidence from experiments with and without primitives suggests that mistakes are corrected we do not know of a paper that tests both environments with the same sample. The closest evidence to compare between the two environments that we found is what we provided in the previous two paragraphs, so from the literature it is not possible to know if in the long run there would be a treatment effect.

A.2 Learning theory in experiments: A brief description of recent related papers

There is a large set of experiments in which feedback of some sort plays a role but where feedback is not part of the central object of study. Meanwhile, the literature that focuses specifically on feedback can perhaps be organized into two groups. The first is the relatively large experimental literature that studies how people use feedback to learn in games, which dates at least back to Harrison and Hirshleifer (1989) and Prasnikar and Roth (1992), and is less directly related to our work in this paper. Models such as reinforcement learning (Erev and Roth (1998), Roth and Erev (1995)), directional learning (Selten and Stoecker 1986), adaptive learning (Cheung and Friedman

[65] See Gal (1996), West and Stanovich (2003), Newell and Rakow (2007), Koehler and James (2009, 2010), James & Koehler (2011).

1997), experience-weighted attraction (Camerer and Hua Ho 1999), and rules-based learning (Stahl 2000) were proposed and tested in this literature. The focus is on what kind of model can rationalize how people learn from feedback, mostly in settings in which taking into account the behavior of other players is crucial. For a detailed survey of the literature, see Part 5 of Dhami (2020).

The second group involves a more recent set of papers that are closer to our paper and focus on evaluating subjects' use of feedback in testing long-run predictions of (behavioral) learning theories. Attention is not on exactly on what model better rationalizes how subjects process the feedback, but on whether long-run choices are consistent with learning-theory predictions.[66] Long-run predictions may differ from Nash equilibria for essentially two reasons. The first case concerns with mistakes that are due to off-path play (i.e. incorrect off-path beliefs that are not corrected via feedback), while the second captures congitive limitations that generate on-path mistakes. We provide examples of both cases next.

As a first example of off-path mistakes leading to long-run behavior that is not part of a Nash equilibrium, consider Fudenberg & Vespa (2019). This paper studies experimentally a signaling game presented in Dekel et al. (2004) in which the first player selects to enter or to stay out and the second player is only asked to make a binary choice (Y or Z) only when the first player selects to enter. Player 1 can have two types (A, B). The game has a unique Nash equilibrium in which player 1 enters and player 2 selects Y. In a first treatment, subjects experience 120 repetitions of this game, each time being randomly matched with another participant, and in each repetition Nature randomly assigns a type to player 1. In this case, self-confirming and Nash equilibria coincide. In a second treatment, types are fixed. A player 1 subject assigned type B may initially believe that player 2 would select Z upon entry, and in such case player 1 type B would want to stay out. If she stays out, she would never collect feedback that challenges such beliefs. It is thus possible in the long run that player 1 type B never enters, so with fixed types there is a self-confirming equilibrium that is not Nash. The experiment in Fudenberg & Vespa (2019) presents data in line with the comparative static.

Cognitive limitations of agents are behind the second case capturing long-run play that deviates from Nash play. For example, the notion of Behavioral Equilibrium (Esponda 2008) captures the long-run behavior of an agent that has difficulties to understand endogenous selection in her feedback. An experimental test of these predictions is studied in Esponda & Vespa (2018). An agent who does not control for selection will have a biased view of the environment. Such biased view would lead to decisions that are suboptimal, but could generate feedback that results in a

[66] A central theoretical reference in this literature is Fudenberg and Levine (1998).

Non-Nash equilibrium. The experimental test consists of comparing choices in a treatment in which feedback involves a selected sample against a treatment in which such selection is not present. The evidence suggests that most subjects do not adjust for selection and end up making suboptimal choices in the long run.

Fudenberg & Peysakhovich (2016) study a version of the classic lemons problem (Akerlof 1970) in which subjects observe 30 rounds of feedback and in which on-path mistakes can arise. The experiment is designed to distinguish between different theoretical notions of behavior that capture cognitive limitations (e.g. cursed equilibrium (Eyster and Rabin 2005) and behavioral equilibrium (Esponda 2008)). The data suggests that subjects give more weight to recent observations (i.e. a recency effect), a feature that was not present in behavioral learning models. Connected to our paper, they also find that providing subjects with a processed summary of the information they have observed helps them make better choices.

Relatedly, Barron et al. (2019) study a situation in which individuals try to learn from observing behavior of others who have faced similar decisions previously. However, information from others involves selection because choices of others are observed conditional on private information. Their experimental paper uses the theoretical selection neglect framework of Jehiel (2018). The paper documents evidence of selection neglect, which is consistent with findings in other papers in this literature. They also document that issues with selections increase when the agents generating the feedback that others use have more private information.[67] In all of the papers in this part of the literature the quality of the feedback depends on subject's choices. A difference with our paper is that in the environments we study the quality of subjects' choices is independent of the quality of the feedback that subjects receive.

[67]There is also a related set of papers that do not focus on feedback per se but that also show that taking selection into account is extremely challenging for many subjects. Prominent recent examples include Enke (2020) and Araujo et al. (2021).

B Details on the experimental design

In this appendix, we summarize our experimental design. For full details on the experimental material, see the Procedures Appendix.

Core treatments

The core treatments consist of nine parts. For expositional purposes, in the main text we grouped the nine parts into four. What we described as the first part in Section 2 corresponds to the BRN task (Part 2 below), and the instructions necessary to introduce the elicitation mechanism (Parts 0 and 1 below). The second part in Section 2 maps to Parts 3 and 4. The third and fourth parts, were introduced in Section 4.5. Specifically, the third part corresponds to Parts 5, 6, 7 and 8. The fourth part includes only Part 9. We now briefly summarize what each of the nine parts achieves.

Part 0

This part uses a simple example to describe the BDM belief elicitation method. Specifically we ask subjects to consider a trivial question: "What is the chance that a fair coin lands Heads vs. Tails?" We ask them to submit an answer to this question (non-incentivized) using a similar 0 to 100 slider as we will use in our main task later. Given a selection in the slider (which is initially blank) the top of the slider indicates the percent chance that the coin lands heads that corresponds to the selection and the bottom of the slider describes the percent chance that the coin lands tail that corresponds to the selection. We then describe, given the BDM mechanism, why it is payoff-maximizing to report their best assessment that the coin will land heads. Given that there is an objective answer to this question, we describe qualitatively why answering 50% is optimal.[68]

Part 1

The aim of this part is to introduce the strategy method. There are two decks of cards, each with 100 cards and cards can be green or blue. One card of the 200 cards is randomly selected and they have to indicate the chance that the selected card is green vs. blue in case it belongs to deck 1, and separately, in case it belongs to deck 2. On the screens subjects are informed of the composition of

[68]The BDM mechanism works in the following manner. After subjects submit a choice $X\%$ that the event at the top of the slider happens, the interface uniformly draws a value between 0 and 100, which we call Y. If $Y \geq X$, the subject wins \$25 with $Y\%$ chance. If $Y < X$, the subject wins \$25 if the event occurs.

each deck before they submit their answers. As the problem in Part 0, there is an objective answer to maximize payoffs in this problem. After they submit their answers, an explanation appears on the screen describing the answers that maximize payoffs. They repeat this problem twice, each time with different compositions of each deck.

Part 2

This section involves the main task. For each possible test result (positive, negative) participants submit the chance that the project is a success vs. a failure. The instructions are presented in Appendix J. This is the only part in the experiment where the instructions to treatment *Primitives* differ from those of *NoPrimitives*.

Part 3

Consists of 99 repetitions of the Part 2 task. The Part 2 task is referred to as round 1 of Part 3, participants get feedback on their round 1 choice and subsequently make 99 additional choices, getting feedback in each round. Feedback is presented round by round on a table, where for each round they learn whether the test was positive or negative and whether the project was a success or a failure.

Part 4

This part consist of 100 additional rounds. It is identical to Part 3, except that subjects make a choice every ten rounds.

Part 5

In this part, we ask subjects to recall the feedback they received on the updating task in the last 200 rounds. Specifically, we ask them to recall the number of rounds in which the four possible types of events were observed: positive signal and success, positive signal and failure, negative signal and success, and negative signal and failure. For payment, the interface selects one of the four entries (with equal chance). The subject earns $25 if the number reported is within plus or minus 5 of the actual number that they experienced.

Part 6

In this part, we confront subjects with the actual data they observed in a conveniently aggregated manner. We present the data in a two-by-two table showing the number of actual rounds in which a specific combination of the signal and state realization was observed. Because it was hard to anticipate what kind of concrete feedback would prompt subjects to revise their incorrect beliefs prior to running the experiment, we proceeded in three steps.

In the first step (Part 6), we present subjects with data from the previous 200 rounds that they experienced. After observing this information, subjects do one more round of the belief elicitation task.

Part 7

In the next step, the interface simulates an additional 800 rounds of signal-state realizations, adds it to the existing 200 rounds, and presents the data in the same table format. Thus, subjects now observe feedback from 1,000 rounds in a table format. After observing this information, subjects do one more round of the belief elicitation task.

Part 8

In the last step, the interface computes the relevant frequencies of the entries presented in the table from the previous step. In particular, conditional on each possible signal (positive or negative), the interface reports the percentage of all 1,000 rounds in which the project was a success vs. failure. After observing this information, subjects have to enter it back themselves (to minimize any chance that they are not reading the data) and subsequently do one more round of the belief elicitation task.

Part 9

In the last part of the experiment, we change the primitives of the belief elicitation task to $p' = .95$ and $q' = .85$. Subjects in both the *Primitives* and *NoPrimitives* treatment are informed of these primitives, and subjects submit beliefs once without the possibility of further feedback.

Survey

At the end of the experiment, we conducted a brief survey consisting of four questions to assess whether the subject had taken a class in probability and/or statistics in college, whether or not their major is STEM related, their gender, and their year of study in college (freshman, sophomore, junior, senior, or graduate student).

Mechanism treatments

Primitives w/ shock

This treatment is identical to the Primitives treatment until the beginning of Part 3. After instructions for Part 3 are read but before they receive feedback, the screen displays a message in case their answers to Part 2 were not correct. Specifically, if only one answer was not correct, they would see the following message "At least one of the answers that you provided in Part 2 is NOT CORRECT." If both answers were incorrect, the screen would shot the following message: "Both answers that you provided in Part 2 are NOT CORRECT."

Subsequently, Parts 3 and 4 proceed as in the Primitives treatment. Subjects then face Parts 5 and 6 as in the core treatments.

Primitives w/ lock* in and *NoPrimitives w/ lock in

These treatments are identical to the Primitives and *NoPrimitives* treatment, respectively until Part 3. At that point and for both treatments, the instructions for Part 3 include the following sentences in the last paragraph: "(...) You will also have a 'lock-in' option. This option enables you to use your current choices for the current round and all future rounds. In other words, if you select this option, you will not need to click through all the remaining rounds; instead you will jump to the end of the experiment. But this also means that you will not be able to modify your choice for future rounds. Note that even if you use the 'lock-in' option to skip to the end of the experiment, you will not be able to leave early. We will pay you only after everybody is done. You will be able to make choices at your own pace in this part. Part 3 will end after you make your choices for all rounds." Given the option to lock in choices, we merged parts 3 and 4 in this treatment. Essentially, subjects were told that in Part 3 they would face additional 199 rounds.

	P	NP	P w/ shock	P w/ lock in	NP w/ lock in	P w/ freq.	NP w/ freq.
Part 0	✓	✓	✓	✓	✓	✓	✓
Part 1	✓	✓	✓	✓	✓	✓	✓
Part 2	P version	NP version	P version	P version	NP version	P version	NP version
Message	No	No	If P2 incorrect	Option to lock in	Option to lock in	No	No
Part 3	By round	By round	By round	By round	By round	Aggregates rounds	Aggregates rounds
Part 4							
Part 5	✓	✓	✓	-	-	✓	✓
Part 6	✓	✓	✓	-	-	-	-
Part 7	✓	✓	-	-	-	-	-
Part 8	✓	✓	-	-	-	-	-
Part 9	✓	✓	-	-	-	-	-
N	64	64	70	74	65	59	59
Location	UCSB	UCSB	UCSD	UCSD	UCSD	UCSB	UCSB

Notes: (i) P for Primitives, NP for No Primitives.
(ii) If P2 incorrect: Subjects who answer incorrectly in Part 2 learn that before starting with Part 3.
(iii) Option to lock in: Subjects learn that they can lock-in their choices in Parts 3 and 4.
(iv) By round: feedback table that reports the signal-state pair outcome round by round.
(v) Aggregate rounds: two-by-two feedback table that aggregates the signal-state pairs across rounds.

Table 2: Summary of BRN Treatments

Primitives w/ freq.* and *NoPrimitives w/ freq.

These treatments are identical to the *Primitives* and *NoPrimitives* treatments, respectively, except that the feedback in Parts 3 and 4 is presented in a two-by-two table showing the number of actual rounds in which a specific combination of the signal and state realization was observed so far.[69]

Voting treatments

We conducted for voting treatments. Participants were recruited from Prolific and there are 130 participants per treatment.[70] These treatments have two parts. Full details of instructions with screenshots are provided in the Procedures Appendix.

Part 1

After reading detailed instructions and questions on the instructions, subjects make the decision for Part 1. How the choice between Option 1 and Option 2 changes across the four treatments is described in Table 3. The problem in Complex Primitives (Voting) is the same as the problem in Primitives (Voting) except that the options are described in a less transparent manner. A similar comment applies to the No Primitives treatments.

After subjects submit their choice for Part 1, we ask them: "How confident do you feel about your choice in Part 1?" This question is unincentivized. Possible answers range from 'Not confident at all' to 'Extremely confident,' with three additional options in between.

Part 2

Part 2 consists of 99 rounds, with the first round providing feedback on the Part 1 choice. This part is identical in all treatments. Subjects observe informative feedback, which is exogenous to their choices, as in the BRN treatments. We implement this by telling subjects that they will receive feedback from a different participant. In odd rounds they receive feedback from a participant who selected Option 1. In even rounds they receive feedback from a participant who selected Option 2. After responding understanding questions, they start Part 2.

[69] These treatments do include Part 5 (which asks subjects to recollect the data), but we did not ask Part 6 as it essentially would have implied a repetition of the last choice they made in Part 4. Due to a software error we did not collect the survey variables at the end of these treatments.

[70] We decided to double the sample size relative to the BRN experiments because research suggests that online participants can be noisier (Gupta et al. 2021)

Treatment	Voting	Complex Voting
Option 1	pays A	pays A if only one vote for it If there are two votes: (i) A if $RN \leq X$ (ii) B if $RN \in \{X+1, ...X+10\}$ (iii) C if $RN > X+10$
Option 2	pays B if $RN \leq X$ pays C if $RN > X$	pays B if $RN \leq X-2$ pays A if $RN \in \{X-1, X\}$ pays C if $RC > X$
Option 1 selected?	If there is at least one vote for Option 1	
Option 2 selected?	If there are two votes for Option 2	
Computer's Vote	Option 2 if $RN > X$	

Notes: (i) RN is a random uniform integer in $\{1, ..., 100\}$. Subjects are told that the computer knows RN.
(ii) In *NoPrimitives (Voting)* and *Complex NoPrimitives (Voting)*, subjects are told that A, B, C and X represent numbers, but that they are not be told what the actual numbers are. We also do not tell them what the computer's strategy is or whether it depends on RN.
(iii) In *Primitives (Voting)* and *Complex Primitives (Voting)* subjects know that $A = 0$, $B = 6$, $C = 10$ and $X = 60$. Subjects also know the computer's strategy.
(iv) Option 1 pays the same in both problems. The computer votes for option 1 when $RN \leq X$. So, if there are two votes for option 1 in complex, it pays A. If there is one vote for option 1 in complex, it pays A. Hence, option 1 in complex pays A.
(v) Option 2 pays the same in both problems. The computer votes for option 2 when $RN > X$. If there are two votes for option 2 (and option 2 is only implemented if there are two votes for it), it pays C in both problems.

Table 3: Summary of Voting Treatments: Part 1

They observe feedback in the form of a table, where for each round they can see the other participant's vote and the other participant's payment. They make a choice for each round and the experiment is over once they make the choice for the last round.

C Additional analysis: Results on *Primitives* vs. *NoPrimitives*

C.1 Treatment differences in rounds 1-200

Statistical analysis on treatment differences

	Conditional on positive signal			Conditional on negative signal			H_0
	P	*NP*	*Diff.*	*P*	*NP*	*Diff.*	$P = NP$
Round 1	31	19	$p < 0.001$	18	35	$p < 0.001$	$p < 0.001$
			$p < 0.001$			$p < 0.001$	$p < 0.001$
Round 50	25	19	$p = 0.041$	15	13	$p = 0.599$	$p = 0.107$
			$p = 0.043$			$p = 0.710$	$p = 0.117$
Round 100	24	18	$p = 0.025$	13	8	$p = 0.045$	$p = 0.011$
			$p = 0.026$			$p = 0.053$	$p = 0.011$
Round 200	21	13	$p = 0.002$	10	7	$p = 0.183$	$p = 0.007$
			$p = 0.002$			$p = 0.203$	$p = 0.008$

Table 4: Average Distance to Bayesian Benchmark in *Primitives* vs. *NoPrimitives*

Notes: *P* and *NP* denote *Primitives* and *NPrimitives*. For each round and each treatment the table reports the average of b_j, where b_j is the absolute value of the distance between the submitted belief and the Bayesian benchmark, that is, $b_j = |B_j - B_j^{Bay}|$. At each given round and for each possible signal, the first p-value of the difference corresponds to the p-value of β_j ($j \in \{Pos, Neg\}$) in the following equation: $b_j = \alpha_j + \beta_j P + \upsilon_j$, where; υ_j is an error term; and P is a dummy that takes value 1 if the variable comes from *Primitives*. The second p-value includes three survey controls in each equation: a dummy for whether the subject has taken a probability class, a dummy for whether the subject is enrolled in a STEM major, and a gender dummy. To obtain p-values, we estimate both equations jointly as a system, using seemingly unrelated regressions. This allows us to allow for a correlation across equation (because for a fixed subjects beliefs can be correlated) but assume independence across subjects. Because the regressions are estimated as a system, we can use a Wald test and evaluate the joint hypothesis that there is no treatment effect (i.e. $\beta_{Pos} = \beta_{Neg} = 0$). The p-value of such test (not including and including survey controls) is reported in last column.

Aggregate measure of partial base-rate neglect

Figure 2 presents average beliefs for different rounds relative to the perfect base-rate neglect and Bayesian benchmarks. An alternative way to present our data and highlight treatment differences is to measure the degree to which responses in aggregate display partial base rate neglect. We use an approach that was introduced by Grether (1980) and since has become standard in empirical work studying updating behavior. This approach does not necessarily have a behavioral interpretation, particularly when applied to beliefs submitted over multiple rounds and to a treatment without primitives, but it does provide an indication of how close beliefs are to the benchmark where subjects know the primitives and can apply Bayes' rule by appropriately weighting the prior and the signal accuracy.

To conduct this analysis, we make use of an implication of Bayes' rule that the posteriors

	Conditional on positive signal			Conditional on negative signal			H_0
	P	*NP*	*Diff.*	*P*	*NP*	*Diff.*	$P = NP$
Round 1	64	60	$p = 0.258$	22	39	$p < 0.001$	$p < 0.001$
			$p = 0.297$			$p < 0.001$	$p < 0.001$
Round 50	57	47	$p = 0.028$	18	16	$p = 0.488$	$p = 0.077$
			$p = 0.028$			$p = 0.579$	$p = 0.080$
Round 100	53	47	$p = 0.159$	16	11	$p = 0.035$	$p = 0.056$
			$p = 0.175$			$p = 0.041$	$p = 0.064$
Round 200	54	46	$p = 0.021$	13	10	$p = 0.112$	$p = 0.049$
			$p = 0.025$			$p = 0.123$	$p = 0.055$

Table 5: Average Beliefs in *Primitives* vs. *NoPrimitives*

Notes: *P* and *NP* denote *Primitives* and *NPrimitives*. For each round and each treatment the table reports the average of b_j, where b_j is the submitted belief, that is, $b_j = B_j$. At each given round and for each possible signal, the first p-value of the difference corresponds to the p-value of β_j ($j \in \{Pos, Neg\}$) in the following equation: $b_j = \alpha_j + \beta_j P + \upsilon_j$, where; υ_j is an error term; and P is a dummy that takes value 1 if the variable comes from *Primitives*. The second p-value includes three survey controls in each equation: a dummy for whether the subject has taken a probability class, a dummy for whether the subject is enrolled in a STEM major, and a gender dummy. To obtain p-values, we estimate both equations jointly as a system, using seemingly unrelated regressions. This allows us to allow for a correlation across equation (because for a fixed subjects beliefs can be correlated) but assume independence across subjects. Because the regressions are estimated as a system, we can use a Wald test and evaluate the joint hypothesis that there is no treatment effect (i.e. $\beta_{Pos} = \beta_{Neg} = 0$). The p-value of such test (not including and including survey controls) is reported in last column.

odds ratio (in log form) can be written as a linear function of the prior odds ratio and the signal likelihood ratio. Specifically, we estimate the following regression for each round of our data: $\ln\left(\frac{B_j}{1-B_j}\right) = \alpha \ln\left(\frac{p}{1-p}\right) + \beta \ln\left(\frac{Q_j}{1-Q_j}\right)$, where for $j = \{\text{Pos, Neg}\}$, $Q_{\text{Pos}} = q$ and $Q_{\text{Neg}} = 1 - q$. The parameter α captures responsiveness to the prior (controlling for its strength), while β captures responsiveness to the signal (controlling for its informational value). This provides us with two benchmarks: $\alpha = \beta = 1$ for a Bayesian, and $\alpha = 0$, $\beta = 1$ for a pBRN agent. Importantly, the estimate on α gives us a continuous measure of the level of partial base rate neglect in the aggregate data.[71]

While there are no significant differences in the estimates of β between treatments (and estimates are relatively close to 1), Figure 10 reveals large differences in the estimates of α.

Consistent with our earlier findings, the estimate of α for both treatments remains substantially below the Bayesian benchmark even after 200 rounds. More importantly, the 200-round estimate of α for treatment *Primitives*, which equals .55, is significantly smaller than that of treatment *NoPrimitives*, which is .82 (p-value 0.001). Table 6 summarizes estimates of α and β at round 200 in all our treatments involving the updating task.

[71] To study treatment differences, we pool data from *Primitives* and *NoPrimitives* allowing for different α and β estimates for the two treatments. Reported significance is with respect to the equivalence of the estimates from the two treatments. We cluster standard errors by subject.

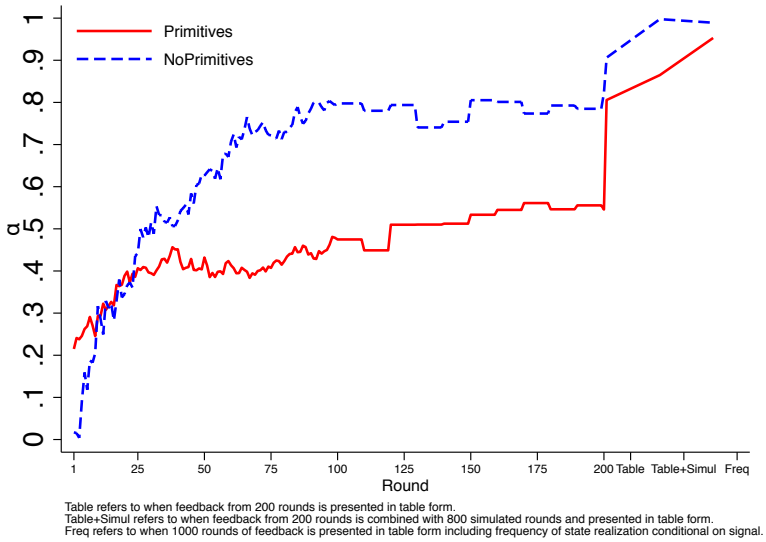


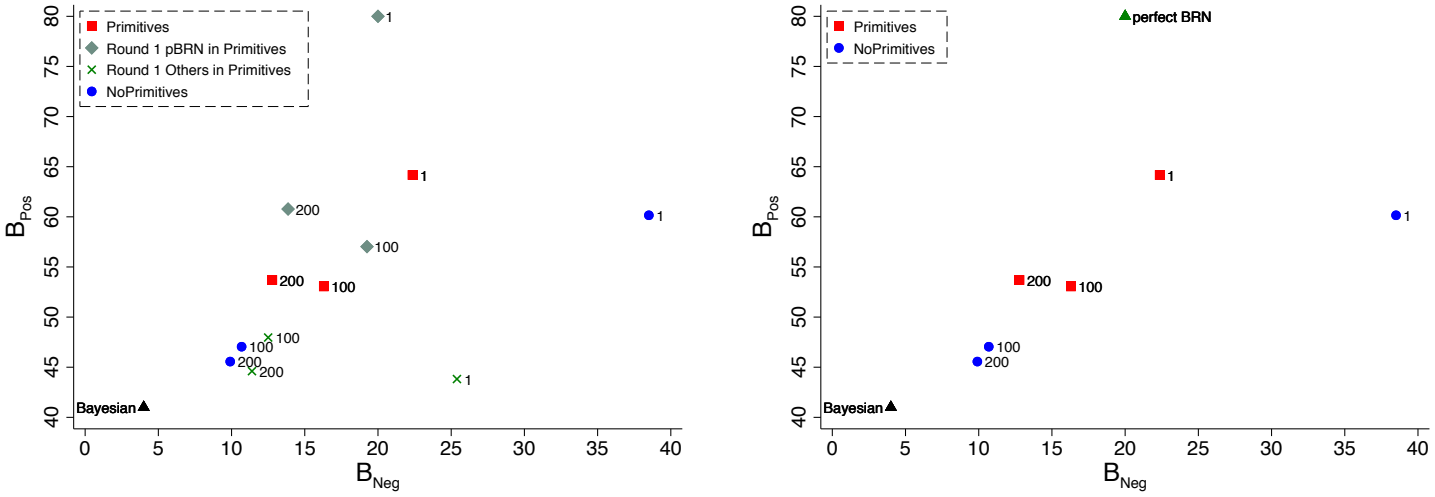
Figure 10: Estimates of α per round by treatment

	Estimates						
	P	*NP*	*Ps*	*Pl*	*NPl*	*Pf*	*NPf*
α	0.55	0.82	0.82	0.49	0.72	0.89	0.99
β	0.87	0.88	0.81	0.77	0.84	0.83	0.99
	Differences						
	P vs. *NP*	*P* vs. *Ps*	*NP* vs. *Ps*	*Pl* vs. *NPl*	*Pf* vs. *NPf*	*Pf* vs. *P*	*NPf* vs. *NP*
α	$p = 0.001$	$p = 0.001$	$p = 0.987$	$p = 0.014$	$p = 0.127$	$p < 0.001$	$p = 0.015$
β	$p = 0.897$	$p = 0.444$	$p = 0.414$	$p = 0.430$	$p = 0.334$	$p = 0.412$	$p = 0.122$

Notes: *P* and *NP*, *Ps*, *Pl*, *NPl*, *Pf*, *NPf*, denote *Primitives* and *NoPrimitives*, *Primitives w/ shock*, *Primitives w/ lock in*, *NoPrimitives w/ lock in*, *Primitives w/ freq*, and *No Primitives w/ freq*. Reported values correspond to the following regression for round 200: $\ln\left(\frac{B_j}{1-B_j}\right) = \alpha \ln\left(\frac{p}{1-p}\right) + \beta \ln\left(\frac{Q_j}{1-Q_j}\right)$, where for $j = \{\text{Pos, Neg}\}$, $Q_{\text{Pos}} = q$ and $Q_{\text{Neg}} = 1 - q$. The parameter α captures responsiveness to the prior (controlling for its strength), while β captures responsiveness to the signal (controlling for its informational value).

Table 6: Estimates from Grether Regressions in Round 200

Behavior of Round 1 pBRN subjects vs. Others in *Primitives*



(a) Decomposition in *Primitives*: Rounds 1, 100 and 200

(b) $B_{Pos} \in [70, 100]$ and $B_{Neg} \in [0, 30]$

Figure 11: Evolution of submitted beliefs by subgroups

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Triangles indicate the Bayesian and the pBRN benchmarks. Squares (Circles) report averages in treatment *Primitives* (*NoPrimitives*). Diamonds indicate averages for R1 pBRN subjects in *Primitives*. Crosses indicate average for R1 other subjects in *Primitives*. The numbers indicate the round for which the averages are taken.

Figure 11a separately follows with diamonds the behavior of Round 1 pBRN subjects. Note that, by definition, all Round 1 pBRN subjects make pBRN choices in round one, so that the starting point for this group is $(B_{Pos}, B_{Neg}) = (80, 20)$. While beliefs for these subjects move towards the Bayesian benchmark with experience, by round 200 beliefs for these subjects are substantially farther away from the Bayesian benchmark relative to the average in *Primitives*. Furthermore, the beliefs of Round 1 pBRN subjects are significantly different from subjects in *NoPrimitives*. This is shown in column (1) of Table 7; for example, there is a significant fifteen percentage-point difference in the average of B_{Pos} between the two groups.[72] Here, we focus on Round 1 pBRN subjects who made pBRN choices in round 1, but may change their behavior as the session evolves. Additionally, it is possible to trace the proportion of subjects in each round who make choices consistent with pBRN. Such evolution is presented in Figure 12.

[72] If we test the joint hypothesis that there are differences in B_{Pos} and B_{Neg}, we obtain p-values of 0.007 and 0.001 in rounds 100 and 200, respectively.

Sample	(1) Round 1 pBRN v. NoPrimitives		(2) Round 1 pBRN v. Round 1 Others		(3) Round 1 Others v. NoPrimitives	
	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}
Round 1	19.8	-18.5	36.2	-5.4	-16.3	-13.1
	(.000)	(.000)	(.000)	(.192)	(.000)	(.002)
Round 100	10.0	8.6	9.1	6.8	0.9	1.8
	(.047)	(.010)	(.157)	(.115)	(.858)	(.486)
Round 200	15.2	3.9	16.2	2.5	-0.9	1.5
	(.000)	(.068)	(.003)	(.279)	(.808)	(.539)
#Obs	100		64		92	

Table 7: Estimation output for subsets of subjects

Notes: The table presents different estimates of γ_{Pos} and γ_{Neg}, where $B_{Pos} = \delta_{Pos} + \gamma_{Pos}P + \upsilon_{Pos}$ and $B_{Neg} = \delta_{Neg} + \gamma_{Neg}P + \upsilon_{Neg}$. Equations are estimated jointly using the seemingly unrelated regressions procedure. In (1) the dummy P takes value 1 if the subject was classified as Round 1 pBRN in *Primitives* and 0 if the subject is in *NoPrimitives*. In (2) the dummy P takes value 1 if the subject was classified as Round 1 pBRN in *Primitives* and 0 if the subject is not classified as Round 1 pBRN in *Primitives* (what we refer to as Round 1 Others in *Primitives*). In (3) dummy P takes value 1 if the subject is classified as 'Round 1 Others in Primitives' and 0 if the subject is in *NoPrimitives*. Between parentheses we report standard errors. Each row constrains the sample to the decision referred to in the first column. The last row indicates the number of observations in each regression.

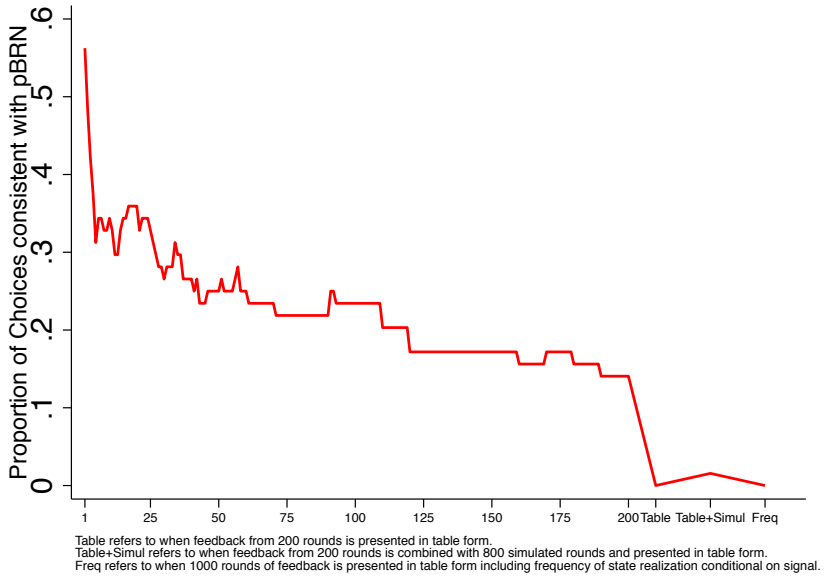


Figure 12: Proportion of choices consistent with pBRN in *Primitives* as the session evolves

In Figure 11b, we demonstrate that these distinct patterns observed for Round 1 pBRN subjects are not due to the fact that they start out in round one with particularly extreme beliefs that are quite far from the Bayesian benchmark. To do so, we study treatment differences focusing on a subset of subjects who start with similar initial beliefs. Specifically, we constrain the sample in both treatments to include only subjects with $B_{Pos} \in [70, 100]$ and $B_{Neg} \in [0, 30]$ in round 1. In

Sample	(1) Round 1 pBRN v. NoPrimitives		(2) $B_{Pos} \geq 70$ $B_{Pos} \leq 30$		(3) Round 1 pBRN v. Round 1 Others		(4) Round 1 Others v. NoPrimitives	
	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}
Round 1	19.8	-18.5	2.0	0.4	36.2	-5.4	-16.3	-13.1
	(.000)	(.000)	(.186)	(.790)	(.000)	(.192)	(.000)	(.002)
Round 100	10.0	8.6	12.5	11.6	9.1	6.8	0.9	1.8
	(.047)	(.010)	(.079)	(.015)	(.157)	(.115)	(.858)	(.486)
Round 200	15.2	3.9	15.5	6.4	16.2	2.5	-0.9	1.5
	(.000)	(.068)	(.012)	(.008)	(.003)	(.279)	(.808)	(.539)
Table -1000- freq	-0.1	3.3	1.4	2.4	0.7	3.6	-0.8	-0.3
	(.930)	(.091)	(.336)	(.483)	(.534)	(.216)	(.577)	(.548)
#Obs	100		60		64		92	

Table 8: Estimation output for subsets of subjects

Notes: The table presents different estimates of γ_{Pos} and γ_{Neg}, where $B_{Pos} = \delta_{Pos} + \gamma_{Pos}\mathrm{P} + \upsilon_{Pos}$ and $B_{Neg} = \delta_{Neg} + \gamma_{Neg}\mathrm{P} + \upsilon_{Neg}$. Equations are estimated jointly using the seemingly unrelated regressions procedure. In (1) the dummy P takes value 1 if the subject was classified as Round 1 pBRN in *Primitives* and 0 if the subject participated in *NoPrimitives*. In (2) P takes value 1 if the subject is in *Primitives* and as 0 if in *NoPrimitives*, but the sample is restricted to subjects who in round 1 submitted beliefs such that: $B_{Pos} \geq 70$ and $B_{Neg} \leq 30$. In (3) the dummy P takes value 1 if the subject was classified as Round 1 pBRN in *Primitives* and 0 if the subject not classified as Round 1 pBRN in *Primitives* (what we refer to as R1 Others in *Primitives*). In (4) dummy P takes value 1 if the subject is classified as 'Round 1 Others in *Primitives*' and 0 if the subject participated in *NoPrimitives*. Between parentheses we report standard errors. Each row constrains the sample to the decision referred to in the first column, where Table-1000-freq refers to the decision after we provide subjects with the relevant frequencies from the 1000-round table. The last row indicates the number of observations in each regression.

Primitives, only Round 1 pBRN subjects are included with this constraint, while in *NoPrimitives* approximately thirty percent of subjects (who likely assigned high informational value to the signal labels) satisfy the constraint. Even within this subset, large treatment differences emerge by round 100, and these differences remain by round 200. Table 8 verifies these patterns statistically.

To provide further evidence that the treatment differences are driven by the Round 1 pBRN subjects, we also separately analyze beliefs of those subjects who are not classified as Round 1 pBRN in *Primitives*. We refer to such subjects as *Round 1 Others*. Average beliefs for these subjects in rounds 1, 100, and 200 are depicted (with crosses) in Figure 11a. At the 100-round and the 200-round marks, average beliefs of Round 1 Others are statistically different from Round 1 pBRN subjects in *Primitives*, but not statistically different from subjects in *NoPrimitives*.[73]

In summary, the decomposition of subjects in *Primitives* depending on their round one choices shows that beliefs of Round 1 pBRN subjects in round 200 are statistically different from other subjects in the same treatment and from subjects in *NoPrimitives*. But such differences are not present between subjects in *NoPrimitives* and subjects in *Primitives* who were not classified as

[73]The p-value of the joint test of $\gamma_{Pos} = \gamma_{Neg} = 0$ by round 200 for the estimates reported in column (3) of Table 7 equals .011, but the same test for estimatates in column (4) delivers a p-value of .760.

Round 1 pBRN, and the beliefs of subjects in these groups are closer to the Bayesian benchmark than the beliefs of Round 1 pBRN subjects in *Primitives*.

Convergence and time

We also use convergence as a measure of when subjects stop responding to data. We code a subject's beliefs to have converged by round t if the subject does not change either belief from round t until round 100.[74] We use $t = 91$ ($t = 96$) to look at the share of subjects whose beliefs converged by the last 10 (5) rounds. We find substantial differences between the treatments. The share of subjects whose beliefs converged by the last 10 rounds is 77 percent in *Primitives* and this share increases to 94 percent when we focus on the last 5 rounds. By contrast, the corresponding values for *NoPrimitives* are only 36 and 47 percent.

Similar patterns are observed with respect to the time that subjects take to make their decisions. The average (median) amount of minutes that subjects in *NoPrimitives* take to complete the first 100 rounds is 15 (12.5), while subjects in *Primitives* take 10.7 (9.2). That is, subjects in *NoPrimitives* take about 30 percent more time relative to subjects in *Primitives*, and the difference is statistically significant (p-value 0.001).

C.2 Treatment differences after round 200

Recollection of feedback

In this part of the experiment, we test how well subjects can recall the feedback they experienced in the rounds 1-200. As explained in Online Appendix B, each subject submits four numbers denoting the number of rounds in which each possible signal-state realization was observed.

A first look at results is presented in Table 9a, which shows the average implied frequency of success conditional on each signal calculated from subjects' recollection of feedback and, in the first row, the actual average frequencies that subjects observed.

We find that frequencies implied by the recollection of feedback are farthest away from the actual frequencies for Round 1 pBRN subjects. Note also that for these subjects the frequencies implied by the recollection of feedback deviate from actual frequencies precisely in the direction of

[74] Recall that rounds 101-200 are introduced as a surprise, so when facing the first 100 rounds subjects did not know that they would receive additional feedback.

Signal was:	Positive	Negative
Actual	*.41*	*.04*
Round 1 pBRN	.54	.15
Round 1 Others	.45	.10
NoPrimitives	.47	.11

(a) Frequency of Success: Actual and inferred from reports

Dep. var.:	(1) $\Delta_{B,F}$	(2) $\Delta_{B,R}$	(3) $\Delta_{R,F}$
$D_{\text{Round 1 pBRN}}$	17.9	12.3	14.3
$D_{\text{Round 1 Others}}$	11.4	9.4	8.1
$D_{\text{NoPrimitives}}$	9.8	10.3	9.6
Hypotheses:			
$D_{\text{Round 1 pBRN}} = D_{\text{Round 1 Others}}$	.006	.262	.021
$D_{\text{Round 1 pBRN}} = D_{\text{NoPrimitives}}$	.000	.333	.033
$D_{\text{Round 1 Others}} = D_{\text{NoPrimitives}}$	.454	.719	.542

(b) Differences between beliefs, reports and feedback across treatments

Table 9: Recollection of feedback

Notes: The right-hand side variable in each regression of panel (b) is indicated on the first row. The right-hand side of each regression includes three dummy variables, each taking value 1 when the subject is in *Primitives* and classified as Round 1 pBRN ($D_{\text{Round 1 pBRN}}$), in *Primitives* and classified as Round 1 Others ($D_{\text{Round 1 Others}}$), or in *NoPrimitives* ($D_{\text{NoPrimitives}}$). Coefficient estimates for the dummy variables are reported in the corresponding row. The p-values associated with the null hypothesis that the coefficient equals zero are are all lower than 0.001 and not reported.

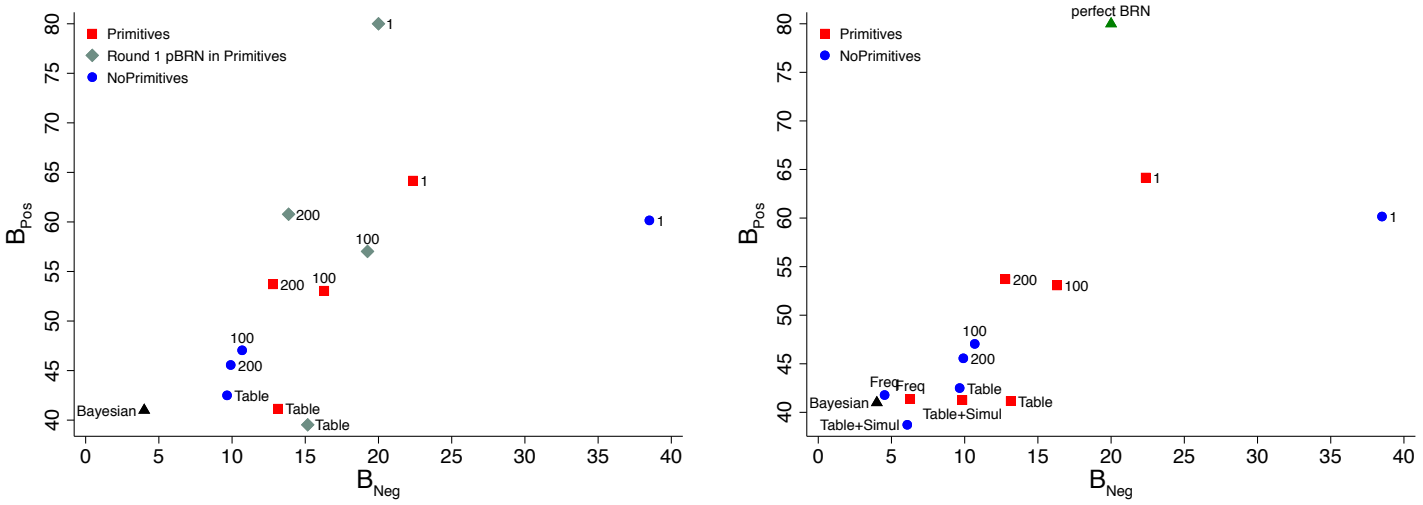
the beliefs they submit.[75]

To study more carefully how well subjects recall feedback and how that connects to the beliefs they submit, in Table 9b we focus on the relationship between three objects: actual realized frequencies (F_j), frequencies implied by recollection of feedback (R_j) and beliefs reported in round 200 (B_j), where $j \in \{\text{Neg}, \text{Pos}\}$.[76] These results can be summarized as follows. (1) We find that frequencies implied by the recollection of feedback, as well as beliefs, to be farthest away from the actual frequencies for Round 1 pBRN subjects at 14.3 and 17.9 percentage points, respectively (see column $\Delta_{R,F}$ and $\Delta_{B,F}$ of Table 9b). While other groups of subjects also have a noisy recollection of the data, the test of hypotheses at the bottom of the table show that such differences are smaller than for Round 1 pBRN subjects. (2) However, there are no statistically significant differences between groups in terms of how far beliefs are from frequencies implied by the recollection of feedback (see column $\Delta_{B,R}$ of Table 9b).

These observations suggest that Round 1 pBRN subjects differ from other subjects in a very specific way. Their beliefs are similarly consistent with their recollection of the data as others, but

[75] This is consistent with subjects using their mental model (due to their limited recollection of past events) to reconstruct what might have happened to them.

[76] We then construct a measure of distance for each subject by computing $\Delta_{x,y} = \frac{|x_{Neg} - y_{Neg}| - |x_{Pos} - y_{Pos}|}{2}$, where x and y represent any two of the objects of interest. We report regressions in which the distance measure is the dependent variable, and the right-hand side includes a dummy variable for each group of subjects (Round 1 pBRN, Round 1 Others and *NoPrimitives*).



(a) Rounds 1, 100, 200 and with summary table

(b) Rounds 1, 100, 200 and with summary tables including simulations

Figure 13: Reported beliefs at different parts of the session

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. 'Table' refers to when subjects are presented with a summary table of the feedback collected in 200 rounds. 'Table + Simul' refers to when the summary table includes 800 additional simulated rounds (for a total of 1000 rounds). 'Freq' refers to when subjects see the table with 1000 rounds of feedback and the relevant frequencies.

they stand out from others in that they have a systematically biased recollection of the data.

Summary tables

In this section we study the effect of showing subjects aggregate data (that they have already experienced) in a summarized table form. As explained in Online Appendix B, we begin by presenting subjects with feedback from rounds 1-200 using a two-by-two table that reports the number of rounds that each of the four combinations of signal-state realizations were observed.[77] We view the provision of the table as an intervention that significantly reduces the attention costs of the subjects.

The main finding is that introducing the table dramatically moves beliefs closer to the Bayesian benchmark in *Primitives*, particularly with respect to B_{Pos}. The movement of average beliefs can

[77]Interventions where subjects are presented with aggregate information is common in the psychology literature. For example, Gigerenzer & Hoffrage (1995) find that providing natural frequencies, as opposed to primitives, reduces, but does not eliminate, base-rate neglect. This literature, however, does not inform on how subjects respond to aggregate information when they are already given the primitives and/or when they have previously experienced the same information directly through natural sampling.

be observed in Figure 13a, in which the average belief for this part of the experiment (denoted 'Table') is shown for different groups. While there is no significant change with respect to B_{Neg}, we observe a downwards adjustment in B_{Pos} of approximately 14 percentage points in treatment *Primitives*.

As explained in Online Appendix B, the part where we provide a summary table is divided into three phases. In the first phase, discussed above, each subject observes a summary table with data from the 200 rounds they experienced. In phases two and three, which we now discuss, subjects observe a summary table from an additional 800 simulated rounds, for a total of 1,000 rounds, and later observe a table with realized frequencies of success and failure conditional on a positive and negative signal. As mentioned earlier, the treatment effect disappears with the first of these interventions. Phases two and three have a small additional impact on beliefs, the main one being that beliefs get closer and closer to the Bayesian benchmark in both treatments. By end of this part, the belief conditional on a positive signal, B_{Pos}, is statistically indistinguishable from the Bayesian belief of 41 percent in both treatments. The belief conditional on a negative signal, B_{Neg}, is statistically different from the Bayesian benchmark of 4 percent in both treatments, but this difference is very small. The findings are presented in the left panel of Figure 14 and Figure 15, which reveal, essentially all subjects in both treatments to report beliefs very close to the Bayesian benchmark by the end of the final phase.

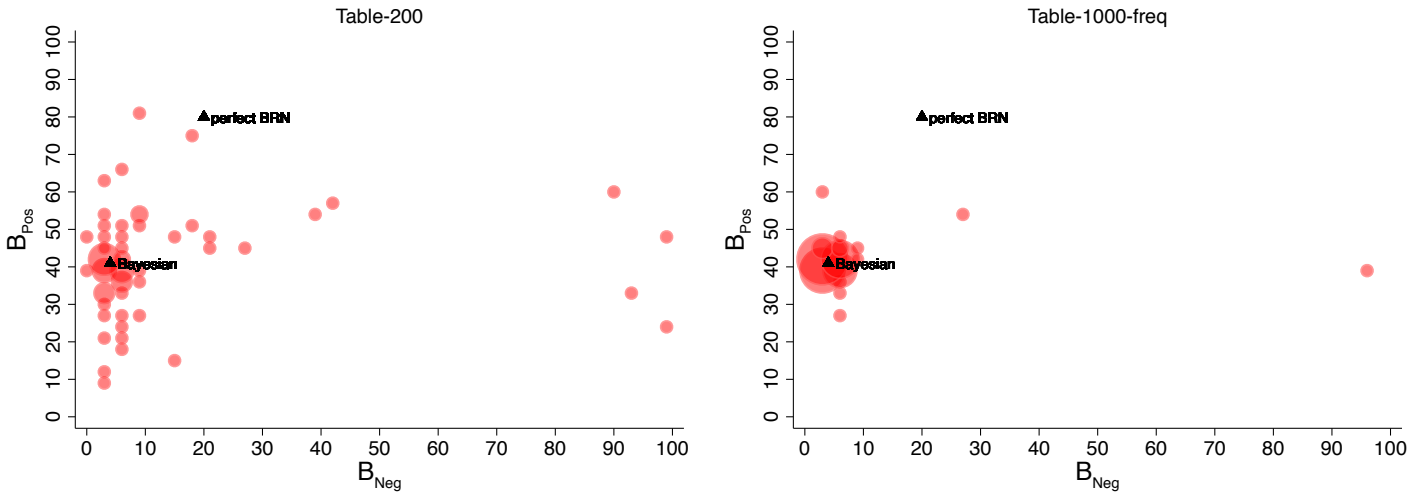


Figure 14: Density plots in the Primitives treatment

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. Table-200 refers to when feedback from 200 rounds is presented in table form. Table-1000-freq refers to when 1000 rounds of feedback is presented in table form including frequency of state realization conditional on signal.

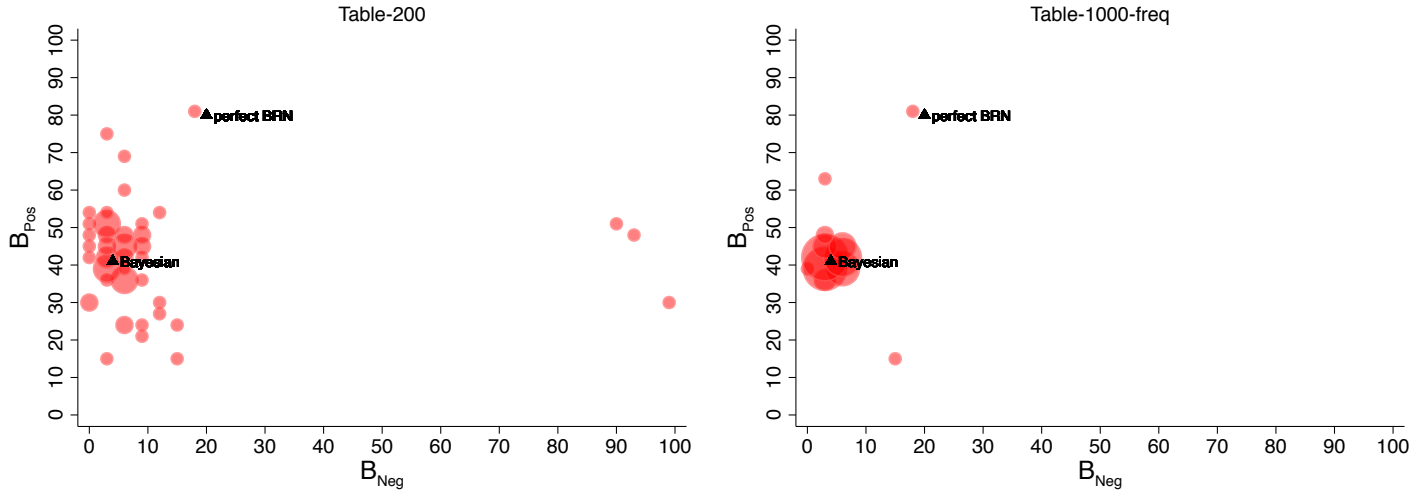


Figure 15: Density plots in the Primitives treatment

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. Table-200 refers to when feedback from 200 rounds is presented in table form. Table-1000-freq refers to when 1000 rounds of feedback is presented in table form including frequency of state realization conditional on signal.

D Additional analysis: Confidence

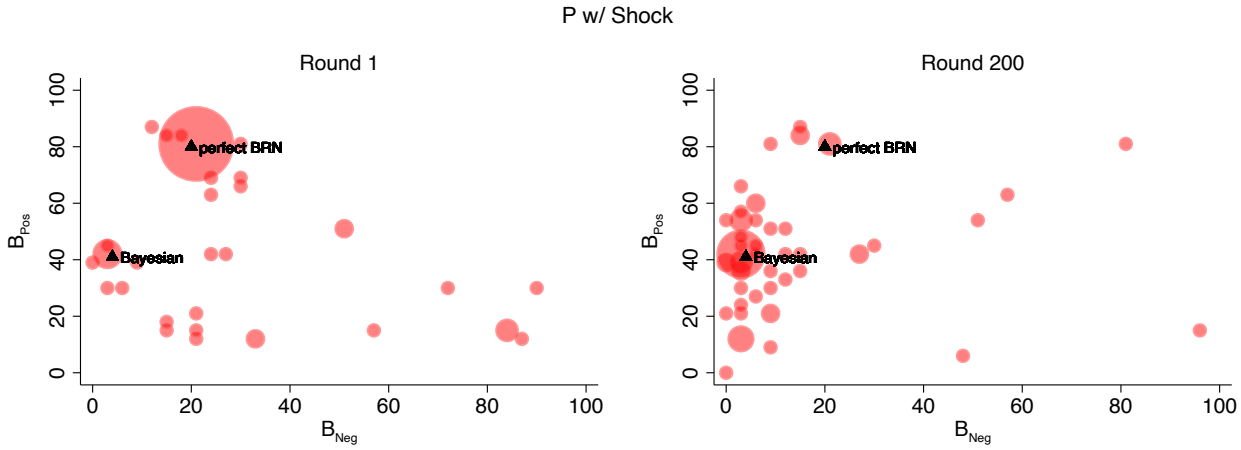


Figure 16: Density Plots for *Primitives w shock*

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.

	Treatment differences		Distance to Bayesian benchmark	
	P+s vs. *P*	*P+s* vs. *NP*	*P+s* vs. *P*	*P+s* vs. *NP*
Round 1	$p = 0.419$	$p < 0.001$	$p = 0.159$	$p < 0.001$
	$p = 0.432$	$p < 0.001$	$p = 0.193$	$p < 0.001$
Round 50	$p = 0.026$	$p = 0.933$	$p = 0.063$	$p = 0.902$
	$p = 0.024$	$p = 0.897$	$p = 0.073$	$p = 0.868$
Round 100	$p = 0.404$	$p = 0.605$	$p = 0.040$	$p = 0.656$
	$p = 0.443$	$p = 0.625$	$p = 0.045$	$p = 0.677$
Round 200	$p = 0.013$	$p = 0.510$	$p = 0.021$	$p = 0.927$
	$p = 0.012$	$p = 0.503$	$p = 0.031$	$p = 0.935$

Table 10: Comparing *Primitives w/ shock* to *Primitives* and *NoPrimitives*

Notes: *P+s*, *P* and *NP* denote *Primitives w/ shock*, *Primitives* and *NoPrimitives*. The first p-value in each comparison results from estimation a system of equations (using seemingly unrelated regressions) for $j \in \{Pos, Neg\}$ given by: $b_j = \alpha_j + \beta_j T + \upsilon_j$, where; υ_j is an error term; and T is a treatment dummy. In columns with the heading 'Treatment differences,' b_j is the submitted belief, that is, $b_j = B_j$. In columns with the heading 'Distance to Bayesian benchmark,' b_j is the absolute value of the distance between the submitted belief and the Bayesian benchmark, that is, $b_j = |B_j - B_j^{Bay}|$. The treatment dummy changes depending on the comparison in the column. For example, in '*P+s vs. P*, it takes value one if the observation comes from *Primitives w/ shock* and zero if it corresponds to *Primitives*. Because the regressions are estimated as a system, we can use a Wald test and evaluate the joint hypothesis that there is no treatment effect (i.e. $\beta_{Pos} = \beta_{Neg} = 0$). Each cell reports the p-value of such test. The second p-value in each comparison results from using the same procedure, but including three right-hand side survey controls: a dummy for whether the subject has taken a probability class, a dummy for whether the subject is enrolled in a STEM major, and a gender dummy.

E Additional analysis: Attentiveness

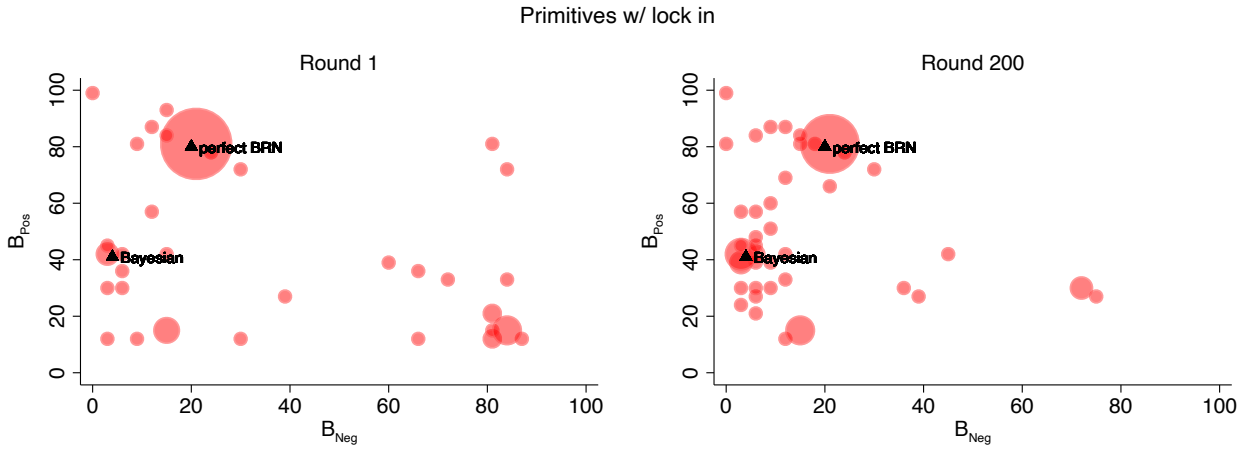


Figure 17: Density Plots for *Primitives w/ lock in*

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

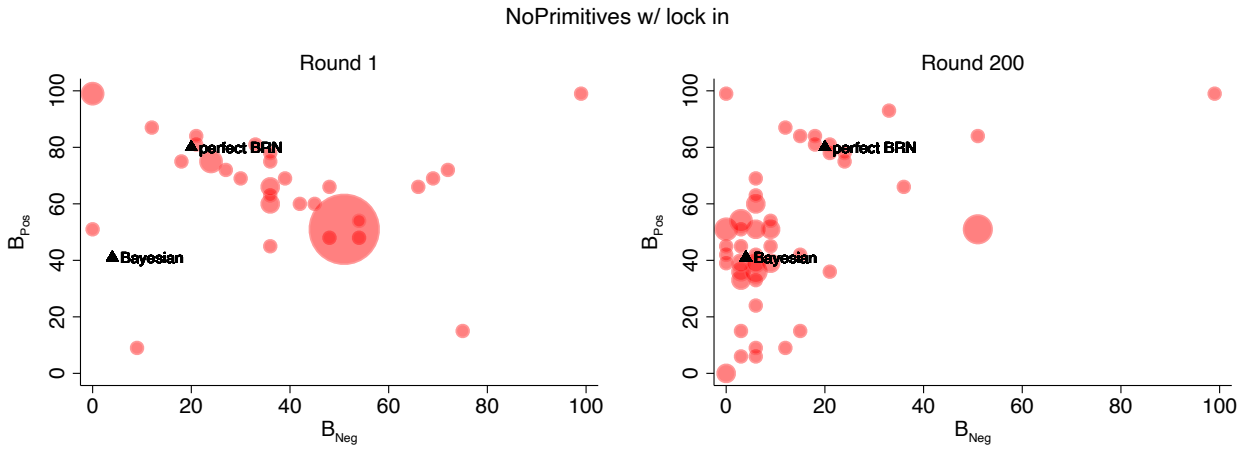


Figure 18: Density Plots for *NoPrimitives w/ lock in*

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

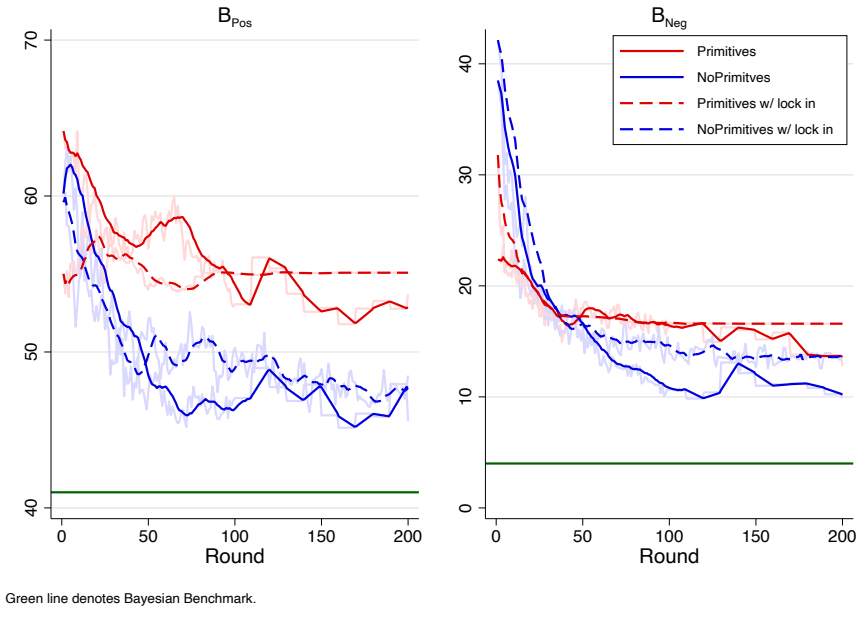


Figure 19: Evolution of Beliefs in Treatments with Lock In

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.

F Additional analysis: Costly attention

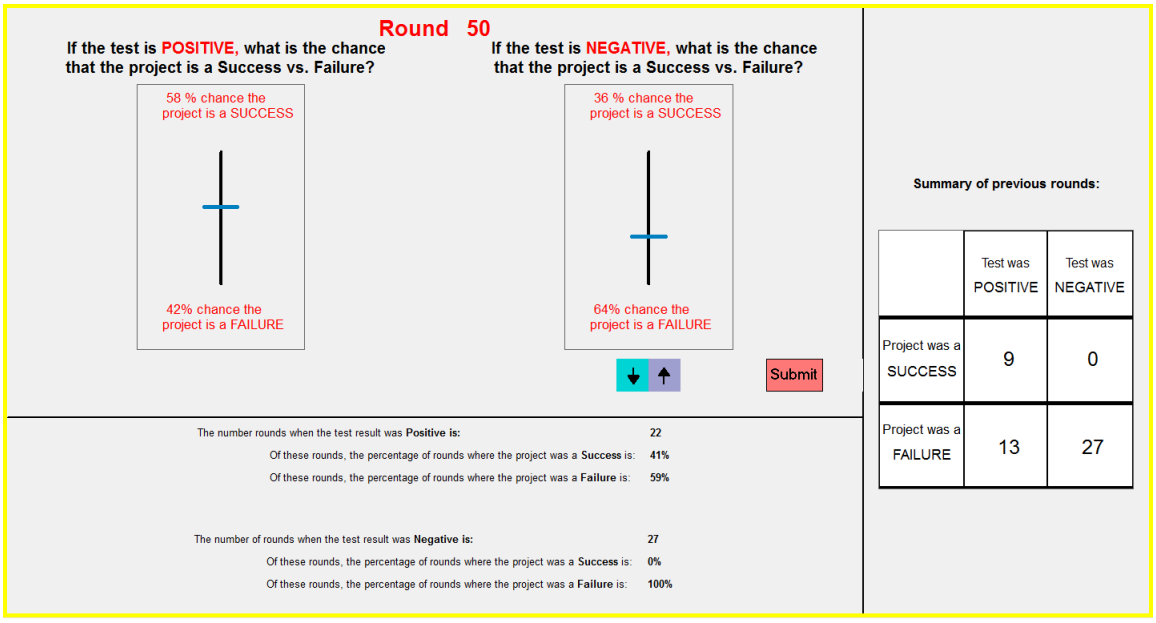


Figure 20: Interface Screenshot of Treatments with Frequencies (Round 50)

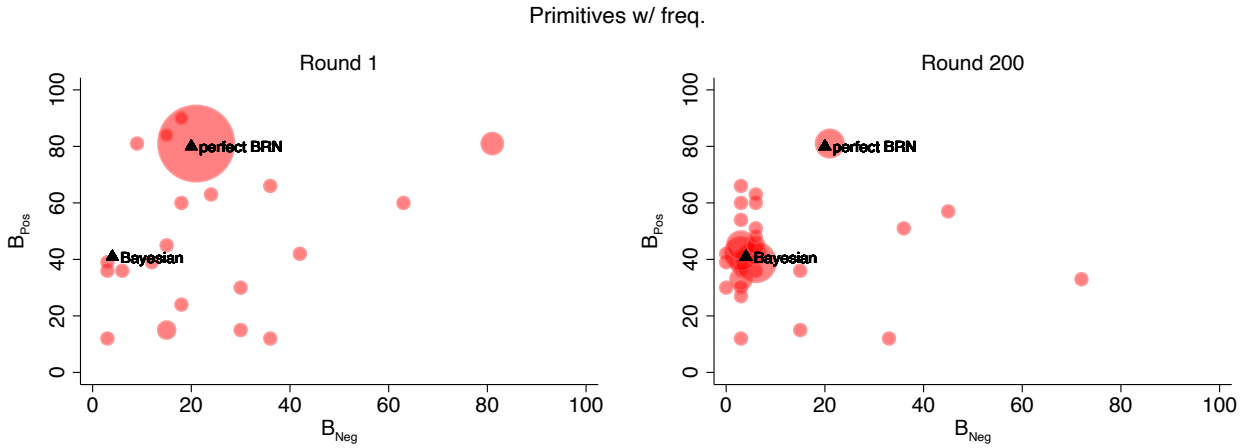


Figure 21: Density Plots for *Primitives w/ freq*

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

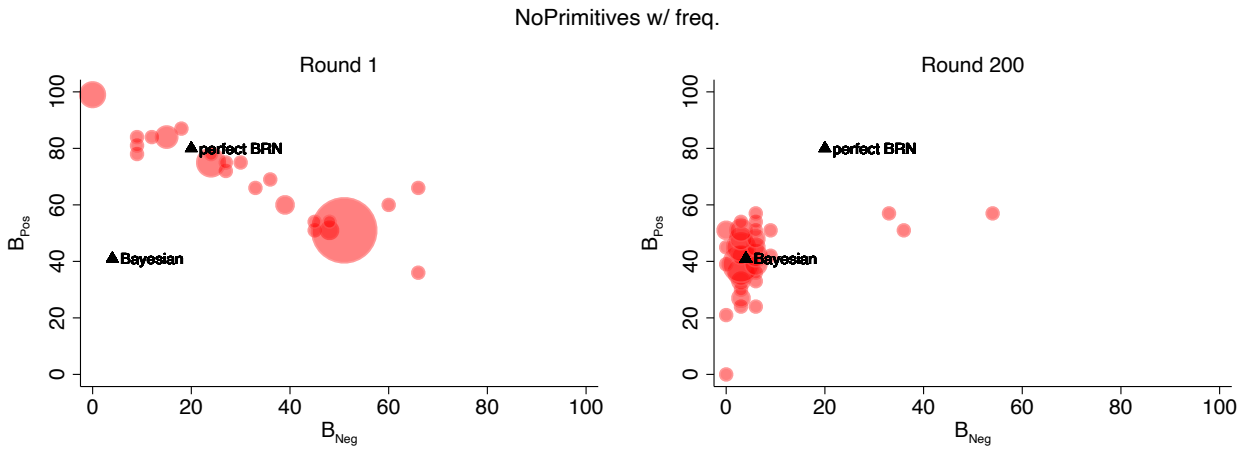


Figure 22: Density Plots for *NoPrimitives w/ freq*

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.

Details on estimation of the learning model on the aggregate level

Estimation of η: We use least squares estimation to find parameters η_j^P and η_j^{NP} for $j \in \{pos, neg\}$ that best describe evolution of beliefs with feedback in *Primitives w/ freq* and *NoPrimitives w/ freq*, respectively.

For each $j \in \{Pos, Neg\}$ and $t \in \{P, NP\}$, we find

$$\underset{\eta_j^t \in \mathbb{R}}{\arg\min} \sum_{r=2,200} \left(\left(\frac{\eta_j^t}{\eta_j^t + n^r} \right) \hat{B}_j^1 + \left(1 - \frac{\eta_j^t}{\eta_j^t + n^r} \right) f^r - \hat{B}_j^r \right)^2,$$

where $\hat{B}_j^t$ is average belief, n^r, average number of observations, and f^r, average empirical frequency at round r.

Estimation of σ: Given estimates for η, we use least squares estimation to find parameters σ_j^P and σ_j^{NP} for $j \in \{Pos, Neg\}$ that best describe evolution of beliefs with feedback in *Primitives* and *NoPrimitives*, respectively.

Taking σ_j^P and σ_j^{NP} as given, for each $j \in \{Pos, Neg\}$ and $t \in \{P, NP\}$, we find

$$\underset{\sigma_j^t \in \mathbb{R}}{\arg\min} \sum_{r=2,200} \left(\left(\frac{\eta_j^t}{\eta_j^t + \sigma_j^t n^r} \right) \hat{B}_j^1 + \left(1 - \frac{\eta_j^t}{\eta_j^t + \sigma_j^t n^r} \right) f^r - \hat{B}_j^r \right)^2,$$

	Treatment differences			Distance to Bayesian benchmark		
Round	*Pf* vs. *NPf*	*Pf* vs. *P*	*NPf* vs. *NP*	*Pf* vs. *NPf*	*Pf* vs. *P*	*NPf* vs. *NP*
1	$p < 0.001$	$p = 0.710$	$p = 0.190$	$p < 0.001$	$p = 0.935$	$p = 0.141$
50	$p = 0.174$	$p = 0.007$	$p = 0.004$	$p = 0.326$	$p < 0.001$	$p < 0.001$
100	$p = 0.272$	$p = 0.005$	$p = 0.058$	$p = 0.394$	$p < 0.001$	$p = 0.001$
200	$p = 0.196$	$p = 0.010$	$p = 0.033$	$p = 0.313$	$p < 0.001$	$p < 0.001$

Notes: *Pf*, *NPf*, *P* and *NP* denote *Primitives w/ freq*, *No Primitives w/ freq*, *Primitives* and *NoPrimitives*. For each cell we estimate a system of equations (using seemingly unrelated regressions) for $j \in \{Pos, Neg\}$ given by: $b_j = \alpha_j + \beta_j T + \upsilon_j$, where; υ_j is an error term; and T is a treatment dummy. In columns with the heading 'Treatment differences,' b_j is the submitted belief, that is, $b_j = B_j$. In columns with the heading 'Distance to Bayesian benchmark,' b_j is the absolute value of the distance between the submitted belief and the Bayesian benchmark, that is, $b_j = |B_j - B_j^{Bay}|$. The treatment dummy changes depending on the comparison in the column. For example, in '*Pf vs. P*, it takes value one if the observation comes from *Primitives w/ freq* and zero if it corresponds to *Primitives*. Because the regressions are estimated as a system, we can use a Wald test and evaluate the joint hypothesis that there is no treatment effect (i.e. $\beta_{Pos} = \beta_{Neg} = 0$). Each cell reports the p-value of such test. Due to a software error, we did not collect survey variables in the frequency treatments.

Table 11: *Primitives w/ freq* and *NoPrimitives w/ freq*

where $\hat{B}_j^t$ is average belief, n^r, average number of observations, and f^r, average empirical frequency at round r.

Estimates on long-run outcomes

The model estimates can also be used to project outcomes for longer horizons than can be observed in our experimental design (beyond 200 rounds). Figure 23 below uses the model to project beliefs for rounds 200 too 1000. While beliefs continue to move towards the Bayesian benchmark in this rage, the qualitative results from the first 200 rounds reported in the paper (particularly the relative comparison of *Primitives* to *NoPrimitives*) remain.

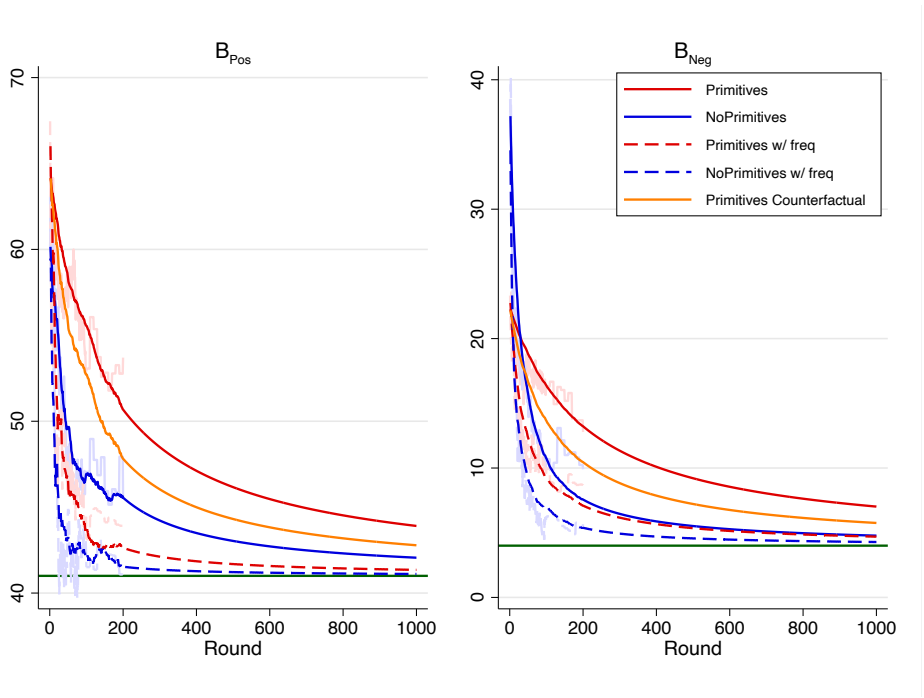


Figure 23: Model Estimates on Evolution of Beliefs for Rounds 1-1000

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). Green line denotes Bayesian benchmark.

Estimating the learning model on the individual level

Here we estimate both ex-ante expected value p_0 and η/σ, which captures relative importance of the prior relative to feedback for each subject in treatments *Primitives*, *NoPrimitives*, *Primitives w/ table* and *NoPrimitives w/ table*. To compute the counterfactual, we need a measure of attentiveness in *NoPrimitives*. We do this by comparing the median estimated value of η/σ in *NoPrimitives* to *NoPrimitives w/ table*. Then we apply this parameter to *Primitives*. We do so by adjusting all individual level estimates from η/σ in *Primitives* by the same ratio so that the median value in this treatment compares to *Primitives w/ table* in the same way as between *NoPrimitives* and *NoPrimitives w/ table*.

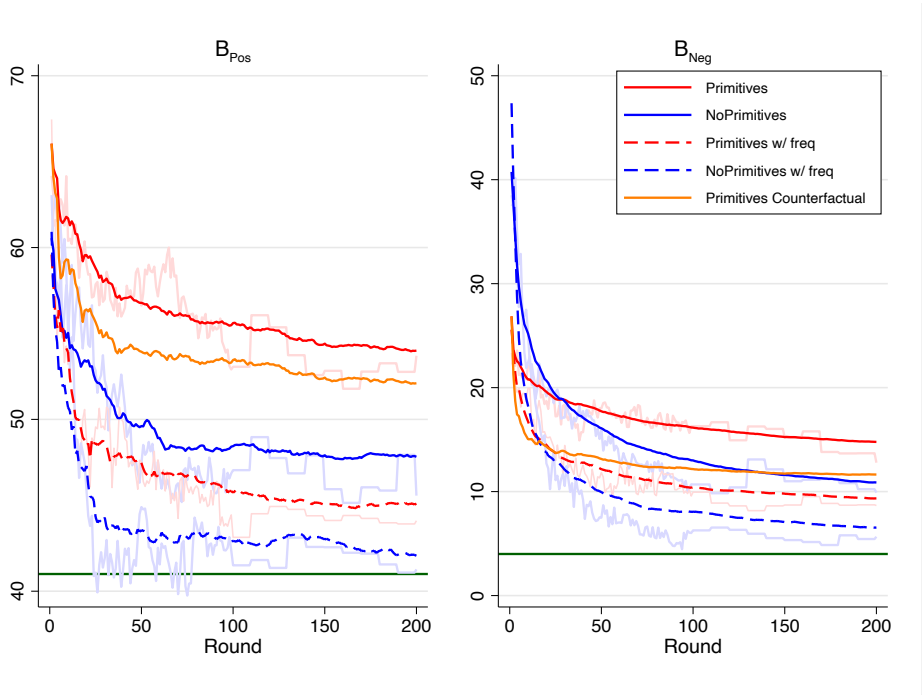


Figure 24: Model Estimates on Evolution of Beliefs Accounting for Heterogeneity

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). Green line denotes Bayesian benchmark.

G Additional analysis: Heterogeneity

In this section, we study the extent to which long-run treatment differences between treatments with and without information on primitives is driven by those subjects who start at the BRN point. We replicate main figures in the main text depicting evolution of beliefs with experience, separately showing beliefs for those subjects who start at the BRN point and others in the same treatment. Tables 12 to 14 provide further information on statistical differences between treatments without primitives and with primitives (separated by subgroups) at important points: rounds 50, 100, and 200, respectively.

Here we summarize key findings:

1. In all cases where there is a long-run difference in beliefs between treatments with and without primitives, this treatment effect is driven by those subjects who start at the BRN point in round one. For evidence on this, see Figures 5 and Figures 26, as well as Tables 12 to 14.

2. In Tables 12 to 14, we go further and look at a subset of subjects in treatments with primitives who start out at the BRN point. Specifically, we separate those subjects who start at the BRN point, but then end up with different beliefs in round 200. Focusing on the contrast between *Primitives* and *NoPrimitives*, we find that beliefs of these subset of subjects in *Primitives* are significantly different from those in *NoPrimitives*. This suggest that the aggregate treatment difference is not driven only by those subjects who never move from the BRN point.

3. Interventions that close or reduce the long-run difference in beliefs between treatments with and without primitives (such as shock to confidence or presentation of feedback as frequency tables) has the largest impact on those subjects who start at the BRN point. For evidence on this, see Figures 25 and Figures 27 particularly, as well as Tables 12 to 14.

4. In treatments with lock-in option, when primitives are provided, those who start at the BRN point lock-in slightly later than others ($p = 0.079$) in the same treatment, but much earlier than those who are not given primitives ($p < 0.001$). However, subjects who start at the BRN point do not keep revising their beliefs for significantly longer than others in the same treatment (but both groups stop revising earlier than those who are not given primitives).[78] This indicates that information on primitives lowers engagement with the data for *all* subjects (both those who start at the BRN point and others). This suggests that subjects classified

[78] This is also the case in other treatments, except in *Primitives w shock* where subjects starting at the BRN point revise their beliefs for longer than others in the same treatment. Furthermore, these pattern do not change when we control for those subjects who are the Bayesian benchmark in round one.

as others in treatments with primitives learn both from data and from primitives. See Tables 15 and 16 for details.

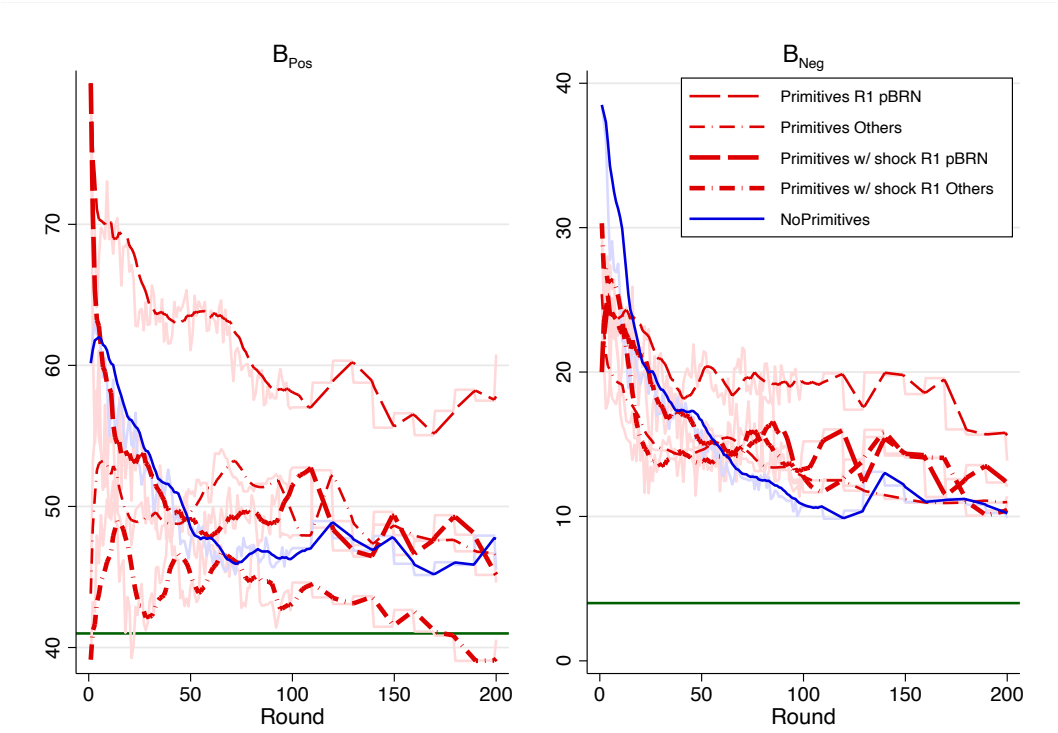


Figure 25: Evolution of Beliefs for R1 pBRN Subjects and Others in *Primitives* and*Primitives w/ shock* vs. *NoPrimitives*

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

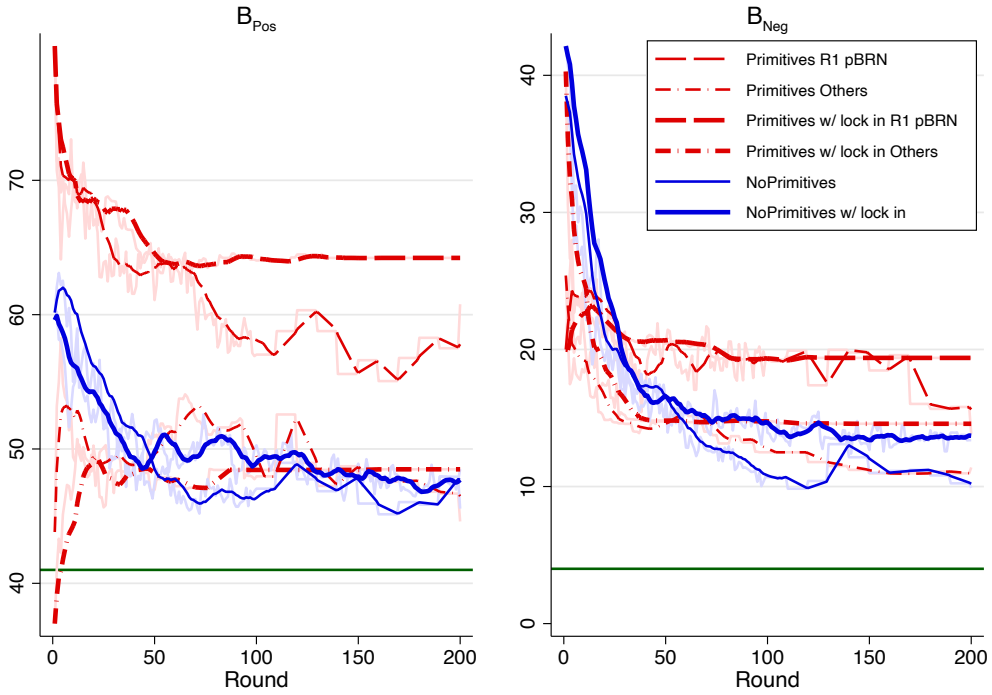


Figure 26: Evolution of Beliefs for R1 pBRN Subjects and Others in *Primitives* and *Primitives w/ lockin* vs. *NoPrimitives*

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

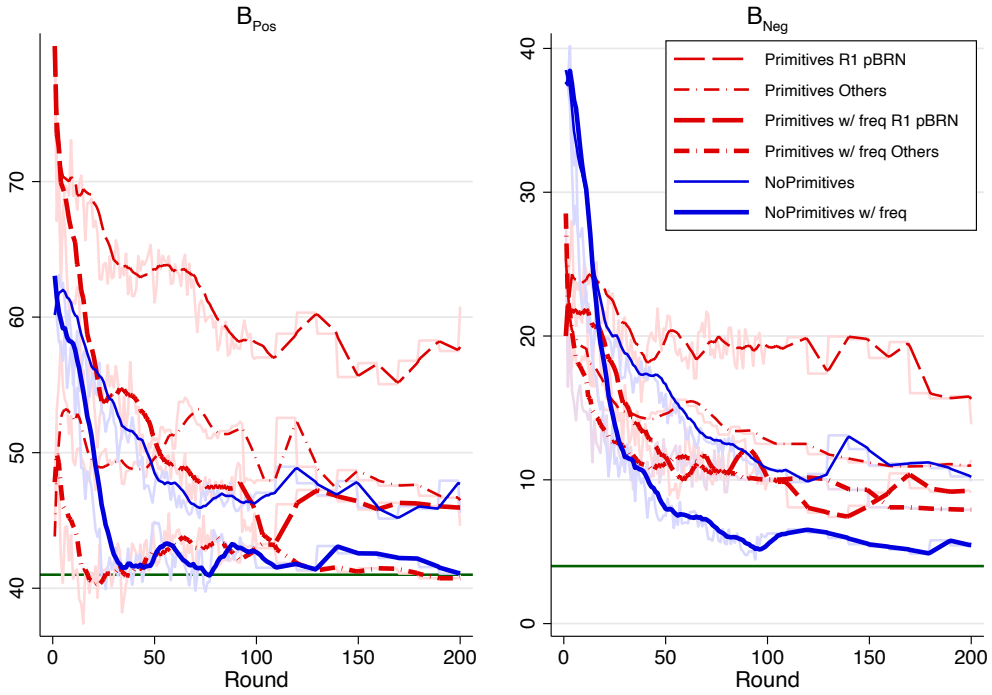


Figure 27: Evolution of Beliefs in *Primitives*, *Primitives w/ freq*, *NoPrimitives* and *NoPrimitives w/ freq*

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

Table 12: Round 50

	Share (%)	B_{Pos}	B_{Neg}		Δ_{Pos}	Δ_{Neg}	
NP		47	15		19	13	
P pBRN in R1	56	61	20	$p = 0.011$	28	17	$p = 003$
P pBRN in R1 but not in R200	44	54	19	$p = 0.078$	23	15	$p = 0.056$
P Others	44	51	15	$p = 0.706$	20	12	$p = 0.859$
NP w/ lockin		52	16		21	14	
P w/ lockin pBRN in R1	42	64	21	$p = 0.050$	27	17	$p = 0.145$
P w/ lockin pBRN in R1 but not in R200	24	49	17	$p = 0.696$	20	13	$p = 0.658$
P w/ lockin Others	58	47	15	$p = 0.543$	20	11	$p = 0.681$
P w/ shock pBRN in R1	49	48	16	$p = 0.995$	20	13	$p = 0.969$
P w/ shock pBRN in R1 but not in R200	47	46	15	$p = 0.995$	18	12	$p = 0.998$
P w/ shock Others	51	44	15	$p = 0.790$	17	11	$p = 0.611$
NP w/ freq		45	7		13	6	
P w/ freq pBRN in R1	61	50	12	$p = 0.126$	19	9	$p = 0.044$
P w/ freq pBRN in R1 but not in R200	53	45	11	$p = 0.513$	13	8	$p = 0.464$
P w/ freq Others	49	43	11	$p = 0.352$	10	8	$p = 0.490$

Notes: P and NP denote *Primitives* and *NPrimitives*. The table reports the average belief (B_{Pos} or B_{Neg}) or average distance to the Bayesian benchmark ($\Delta_{Pos} = |B_{Pos} - B^{Bay}_{Pos}|$ and $\Delta_{Neg} = |B_{Neg} - B^{Bay}_{Neg}|$). The first p-value reports whether beliefs in selected group in P are different from closest NP treatment. The second p-value reports whether distance to Bayesian benchmark is different relative to closest NP treatment. For details of regressions see Table 4. For each group, the closest NP treatment is listed right above (except for treatment with shock where the original NP treatment is considered).

Table 13: Round 100

	Share (%)	B_{Pos}	B_{Neg}		Δ_{Pos}	Δ_{Neg}	
NP		47	11		17	8	
P pBRN in R1	56	57	19	$p = 0.007$	26	16	$p < 0.001$
P pBRN in R1 but not in R200	44	49	16	$p = 0.078$	22	13	$p = 0.021$
P Others	44	48	13	$p = 0.784$	20	10	$p = 0.664$
NP w/ lockin		50	14		19	11	
P w/ lockin pBRN in R1	42	64	19	$p = 0.015$	27	16	$p = 0.050$
P w/ lockin pBRN in R1 but not in R200	24	50	16	$p = 0.660$	20	13	$p = 0.658$
P w/ lockin Others	58	48	15	$p = 0.853$	20	11	$p = 0.886$
P w/ shock pBRN in R1	49	45	12	$p = 0.196$	18	11	$p = 0.514$
P w/ shock pBRN in R1 but not in R200	47	48	13	$p = 0.258$	16	10	$p = 0.546$
P w/ shock Others	51	45	12	$p = 0.749$	15	9	$p = 0.721$
NP w/ freq		42	6		10	5	
P w/ freq pBRN in R1	61	43	10	$p = 0.391$	16	7	$p = 0.077$
P w/ freq pBRN in R1 but not in R200	53	40	9	$p = 0.456$	11	7	$p = 0.542$
P w/ freq Others	49	43	10	$p = 0.424$	8	7	$p = 0.509$

Notes: P and NP denote *Primitives* and *NPrimitives*. The table reports the average belief (B_{Pos} or B_{Neg}) or average distance to the Bayesian benchmark ($\Delta_{Pos} = |B_{Pos} - B^{Bay}_{Pos}|$ and $\Delta_{Neg} = |B_{Neg} - B^{Bay}_{Neg}|$). The first p-value reports whether beliefs in selected group in P are different from closest NP treatment. The second p-value reports whether distance to Bayesian benchmark is different relative to closest NP treatment. For details of regressions see Table 4. For each group, the closest NP treatment is listed right above (except for treatment with shock where the original NP treatment is considered).

Table 14: Round 200

	Share (%)	B_{Pos}	B_{Neg}		Δ_{Pos}	Δ_{Neg}	
NP		46	10		14	7	
P pBRN in R1	56	61	14	$p = 0.001$	25	11	$p < 0.001$
P pBRN in R1 but not in R200	44	50	12	$p = 0.073$	18	9	$p = 0.022$
P Others	44	45	11	$p = 0.760$	15	8	$p = 0.762$
NP w/ lockin		48	13		18	11	
P w/ lockin pBRN in R1	42	64	19	$p = 0.004$	27	16	$p = 0.027$
P w/ lockin pBRN in R1 but not in R200	24	50	16	$p = 0.497$	20	12	$p = 0.620$
P w/ lockin Others	58	48	15	$p = 0.927$	20	11	$p = 0.834$
P w/ shock pBRN in R1	49	45	12	$p = 0.758$	16	9	$p = 0.717$
P w/ shock pBRN in R1 but not in R200	47	42	11	$p = 0.764$	14	8	$p = 0.852$
P w/ shock Others	51	41	11	$p = 0.229$	12	8	$p = 0.792$
NP w/ freq		41	6		7	3	
P w/ freq pBRN in R1	61	46	9	$p = 0.116$	12	6	$p = 0.071$
P w/ freq pBRN in R1 but not in R200	53	41	8	$p = 0.691$	7	5	$p = 0.789$
P w/ freq Others	49	41	8	$p = 0.550$	6	5	$p = 0.431$

Notes: *P* and *NP* denote *Primitives* and *NPrimitives*. The table reports the average belief (B_{Pos} or B_{Neg}) or average distance to the Bayesian benchmark ($\Delta_{Pos} = |B_{Pos} - B_{Pos}^{Bay}|$ and $\Delta_{Neg} = |B_{Neg} - B_{Neg}^{Bay}|$). The first p-value reports whether beliefs in selected group in P are different from closest NP treatment. The second p-value reports whether distance to Bayesian benchmark is different relative to closest NP treatment. For details of regressions see Table 4. For each group, the closest NP treatment is listed right above (except for treatment with shock where the original NP treatment is considered).

Table 15: Round of Last Revision in Beliefs (OLS)

	(1)	(2)	(3)	(4)
Primitives	-42.84***	-90.99***		-42.79***
	(13.15)	(10.31)		(11.87)
R1 pBRN	-16.43	11.53	43.35**	9.205
	(14.62)	(12.36)	(17.87)	(12.89)
Constant	175.5***	113.1***	90.50***	191.5***
	(7.254)	(6.505)	(12.46)	(6.286)
Observations	128	139	70	118

Standard errors in parentheses.

***1%, **5%, *10% significance.

(1): Data from Primitives and NoPrimitives.

(2): Data from Primitives w/ lockin and NoPrimitives w/ lockin.

(3): Data from Primitives w/ shock.

(4): Data from Primitives w/ freq and NoPrimitives w/ freq.

Table 16: Round of Lock-in Decision (OLS)

	Round of Lock-in
Primitives	-90.09***
	(11.48)
R1 pBRN	24.32*
	(13.76)
Constant	124.5***
	(7.245)
Observations	139

Standard errors in parentheses.

***1%, **5%, *10% significance.

Data from Primitives w/ lockin and NoPrimitives w/ lockin..

H Additional analysis: Transfer learning

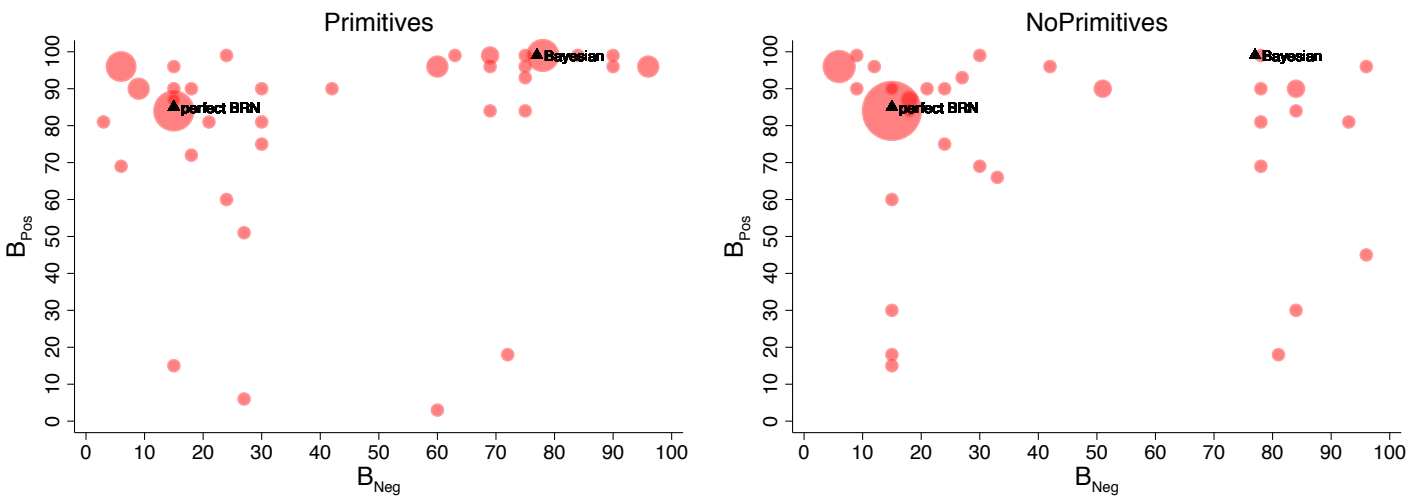


Figure 28: Transfer Learning: Density Plots in Final Round with New Primitives

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.

I Additional Analysis: Evidence beyond the updating problem

In the text we focus on the proportion of choices that are correct in the last round and compare it across treatments. A limitation of this exercise is that it is does not measure convergence. It is possible that subjects making an optimal choice in the last round are still unsettled in their choice and just happened to make an optimal choice at that point. Here we provide an alternative presentation that controls for convergence.

As a reference we will say that a subject converged to the correct choice if the participant made such choice in all the last five rounds. Figures 29 and 30 provide this information. In addition, for each round t, the figures depict for each treatment the proportion of subjects who selected optimally from that round onward.

Consider Figures 29 first. The proportion of subjects who choose correctly from round one onward (i.e. in all rounds) in the *Primitives (Voting)* treatment is approximately 18 percent. These are subjects who very likely identify that there is a dominant vote from the instructions and, hence, have nothing to learn. In *NoPrimitives (Voting)*, identifying the optimal vote from the instructions is not possible and, accordingly, the proportion of subjects selecting consistently in all rounds is lower, at close to ten percent. However, there is substantial learning in *NoPrimitives (Voting)*. In the last five rounds the difference between treatments is 21.5 percentage points, which is significant (p-value <0.001)[79]. The same type of exercise can be done with a less strict consistency condition on optimality, by relaxing the demand that subjects make no mistakes from round t onward. For example, it is possible to construct the same figure demanding that z percent of choices from round t onward are optimal. While such analysis changes the levels, the treatment effects remain the same for values of $z \in \{70, 75, 80, 85, 90, 95\}$.

Figure 30 provides the same comparison but for "Complex" treatments. In this case, there is little to no difference throughout the session. The last-round proportion of subjects behaving optimally is slightly higher in the environment with no primitives but the difference is not significant (p-value 0.537).

The figures also suggest that it is more demanding to learn from feedback in the Complex environment, even though the actual feedback that subjects receive is structurally identical. To see this, notice that the proportion of subjects behaving optimally in the last five rounds of *NoPrimitives (Voting)* is approximately ten percentage points higher than in *Complex NoPrimitives (Voting)*.

[79] We test the null hypothesis of no difference by running a regression in which the proportion of subjects making optimal choices in the last round is on the left-hand side and the right-hand side includes a treatment dummy.

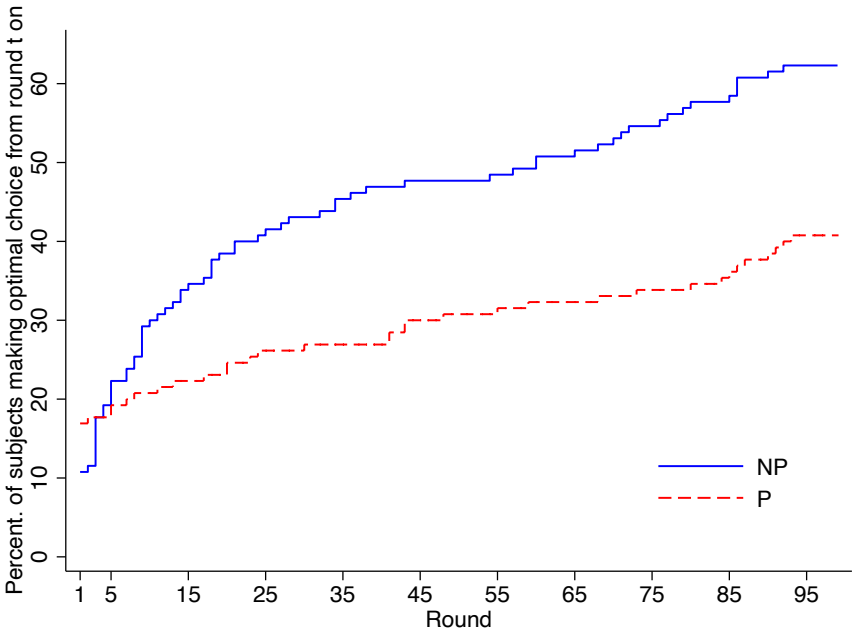


Figure 29: Subjects making optimal choices in *Primitives (Voting)* and *NoPrimitives (Voting)*

Notes: For each treatment the figure reports the percentage of subjects choosing optimally from round t onward.

This suggests that even if the data is of the same quality, having more involved instructions to begin with may make it more difficult for subjects to learn from feedback. It also suggests that the difference between *Primitives (Voting)* and *Complex Primitives (Voting)* may understimate the real difference given that learning in the Complex setting is more challenging. The difference in the last round from comparing these two treatments results in approximately 10 percentage points more subjects behaving optimally in *Complex Primitives (Voting)* than in *Primitives (Voting)*. We leave it for future research to study how learning in settings where options are more difficult to parse to begin with might affect long-run learning.

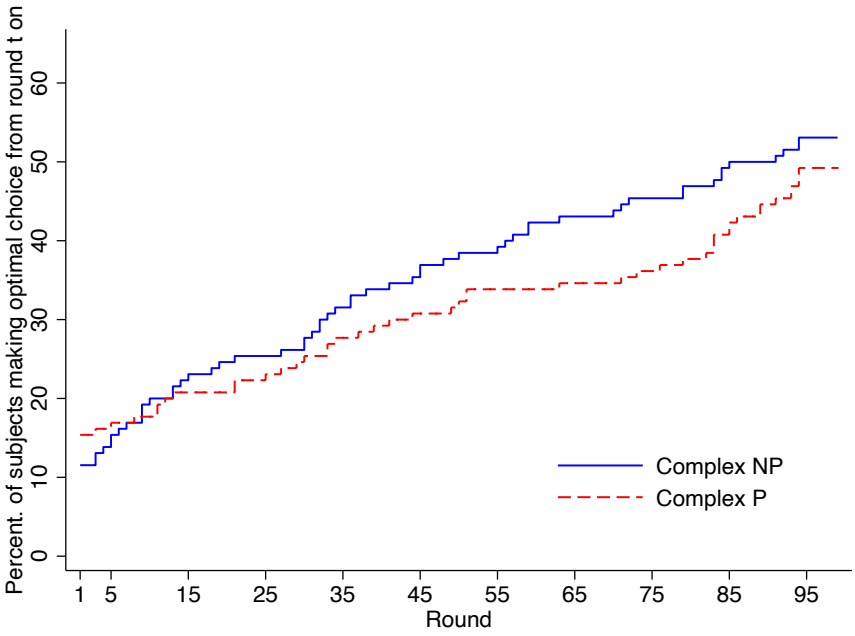


Figure 30: Subjects making optimal choices in *Complex Primitives (Voting)* and *Complex NoPrimitives (Voting)*

Notes: For each treatment the figure reports the percentage of subjects choosing optimally from round t onward.

J Experimental instructions

Full details on our implementation are provided in the Procedures Appendix. In the instructions to the subjects part 2 refers to round 1 as described in the paper. For a more direct access to the crucial differences between treatments in this section, we include the instructions that were presented to subjects on the main updating task (round 1) and how the two treatments (*Primitives* and *NoPrimitives*) differ in this respect. The sections of the instructions that differ by treatment are highlighted between brackets [].

Round 1 Instructions:

There is a total of 100 projects, and one of these projects will be randomly selected (with all projects having an equal chance of being selected).

[*Primitives*: Of the 100 projects, there are 15 projects that are successes and 85 projects that are failures.]

[*NoPrimitives*: Of the 100 projects, a certain number of them are successes and the remaining ones are failures. We will not tell you how many of them are successes and how many are failures.]

Your task is to assess the chance that the project that was randomly selected is a Success vs. Failure.

To aid your assessment, the computer will run a test on the selected project.

[*Primitives*: The test result can be either Positive or Negative and has a reliability of 80%.]

[*NoPrimitives*: The test result can be either Positive or Negative and has a reliability of R%.]

That means that:

[*Primitives*:

- If the project is a Success, the test result will be Positive with 80% chance and the test result will be Negative with 20% chance.
- If the project is a Failure, the test result will be Negative with 80% chance and the test result will be Positive with 20% chance.]

[*NoPrimitives*:

- If the project is a Success, the test result will be Positive with R% chance and the test result will be Negative with (100-R)% chance.
- If the project is a Failure, the test result will be Negative with R% chance and the test result will be Positive with (100-R)% chance.

The reliability R is a specific number between 0 and 100, but we will not tell you this number.]

We will ask you to submit two assessments:

- If the test is Positive, what is the chance that the project is a Success vs. Failure?
- If the test is Negative, what is the chance that the project is a Success vs. Failure?

For each possible test result (Positive and Negative), you will select a point that indicates the chance that the randomly selected project is a Success vs. Failure given the test result. [*NoPrimitives*: Clearly, you are not given enough information to make an informed decision. Please go ahead and take a guess.]

If this part is selected for payment, the interface will first randomly select a project. It will then conduct a test, as described above. If the test result is Positive, we will use your submitted choice for the case where the test is Positive and pay you as explained in the instruction period. If the test result is Negative, we will use your submitted choice for the case where the test is Negative and pay you as explained in the instruction period. The important thing to remember is that to maximize your payment you should give us your best assessment of the chance that the project is a Success vs. Failure given the test result.

Round 1 screenshot (part 2 in instructions):

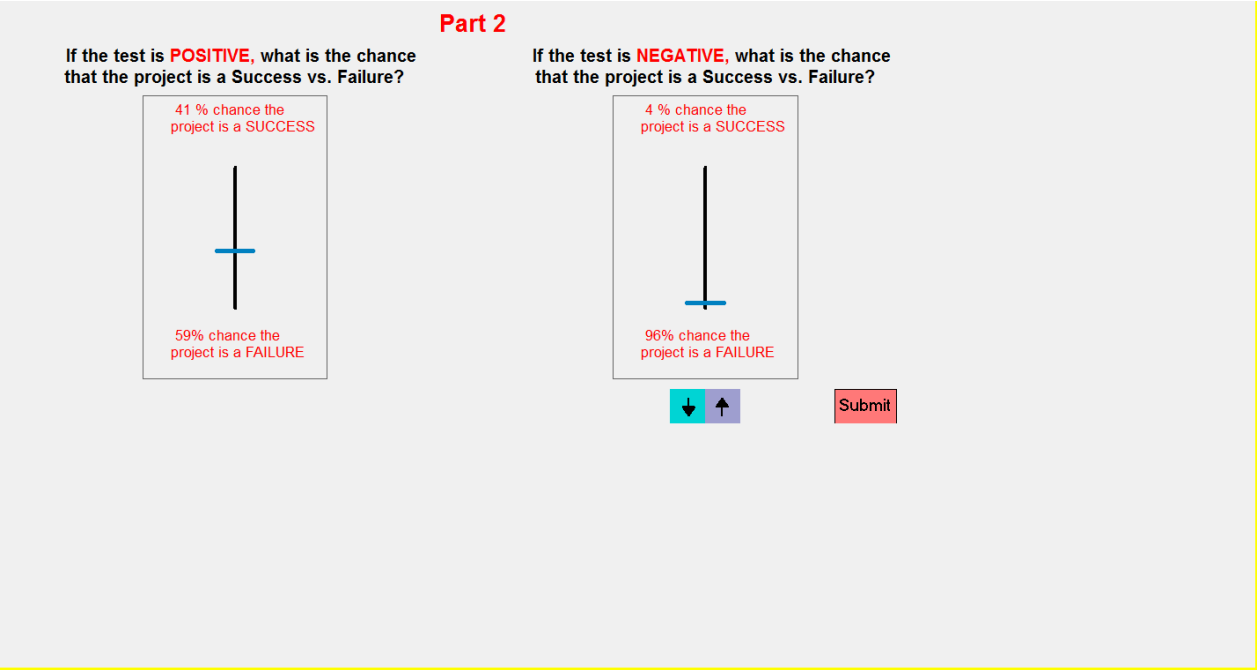


Figure 31: Interface screenshots for round 1 (presented as part 2 to subjects)

Appendix References

Akerlof, Geokge A, "The Market for Lemons: Quality uncertainty and the market mechanism," *The Quarterly Journal of Economics*, 1970, *84* (3), 488–500.

Araujo, Felipe A, Stephanie W Wang, and Alistair J Wilson, *American Economic Journal: Microeconomics*, 2021, *13* (4), 1–22.

Barbey, Aron K and Steven A Sloman, "Base-rate respect: From ecological rationality to dual processes," *Behavioral and Brain Sciences*, 2007, *30* (3), 241–254.

Barron, Kai, Steffen Huck, and Philippe Jehiel, "Everyday econometricians: Selection neglect and overoptimism when learning from others," *Working Paper*, 2019.

Camerer, Colin and Teck Hua Ho, "Experience-weighted attraction learning in normal form games," *Econometrica*, 1999, *67* (4), 827–874.

Cheung, Yin-Wong and Daniel Friedman, "Individual learning in normal form games: Some laboratory results," *Games and economic behavior*, 1997, *19* (1), 46–76.

Christensen-Szalanski, Jay JJ and Lee Roy Beach, "Experience and the base-rate fallacy," *Organizational Behavior and Human Performance*, 1982, *29* (2), 270–278.

Cosmides, Leda and John Tooby, "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty," *cognition*, 1996, *58* (1), 1–73.

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

Dhami, Sanjit, *The Foundations of Behavioral Economic Analysis: Volume VII: Further Topics in Behavioral Economics*, Vol. 7, Oxford University Press, USA, 2020.

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

Erev, Ido and Alvin E Roth, "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American economic review*, 1998, pp. 848–881.

_ **and Ernan Haruvy**, "Learning and the economics of small decisions," *in The handbook of experimental economics*, 2013, *2*, 638–700.

Esponda, I., "Behavioral equilibrium in economies with adverse selection," *The American Economic Review*, 2008, *98* (4), 1269–1291.

Esponda, Ignacio and Emanuel Vespa, "Endogenous sample selection: A laboratory study," *Quantitative Economics*, 2018, *9* (1), 183–216.

Eyster, Erik and Matthew Rabin, "Cursed equilibrium," *Econometrica*, 2005, *73* (5), 1623–1672.

Fantino, Edmund and Anton Navarro, "Description–experience gaps: Assessments in other choice paradigms," *Journal of Behavioral Decision Making*, 2012, *25* (3), 303–314.

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

_ **and David K Levine**, *The theory of learning in games*, Vol. 2, MIT press, 1998.

_ **and Emanuel Vespa**, "Learning Theory and Heterogeneous Play in a Signaling-Game Experiment," *American Economic Journal: Microeconomics*, 2019, *11* (4), 186–215.

Gal, Iddo, "Understanding repeated simple choices," *Thinking & Reasoning*, 1996, *2* (1), 81–98.

Gigerenzer, Gerd, "How to make cognitive illusions disappear: Beyond "heuristics and biases"," *European review of social psychology*, 1991, *2* (1), 83–115.

_ **and Ulrich Hoffrage**, "How to improve Bayesian reasoning without instruction: frequency formats.," *Psychological review*, 1995, *102* (4), 684.

Goodie, Adam S and Edmund Fantino, "Learning to commit or avoid the base-rate error," *Nature*, 1996, *380* (6571), 247.

_ **and** _ , "What does and does not alleviate base-rate neglect under direct experience," *Journal of Behavioral Decision Making*, 1999, *12* (4), 307–335.

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

_ , "Testing Bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior & Organization*, 1992, *17* (1), 31–57.

Griffin, Dale and Amos Tversky, "The weighing of evidence and the determinants of confidence," *Cognitive psychology*, 1992, *24* (3), 411–435.

Gupta, Neeraja, Luca Rigotti, and Alistair Wilson, "The Experimenters' Dilemma: Inferential Preferences over Populations," *Working Paper*, 2021.

Harrison, Glenn W and Jack Hirshleifer, "An experimental evaluation of weakest link/best shot models of public goods," *Journal of Political Economy*, 1989, *97* (1), 201–225.

Jehiel, Philippe, "Investment strategy and selection bias: An equilibrium perspective on overoptimism," *American Economic Review*, 2018, *108* (6), 1582–97.

Kahneman, Daniel and Amos Tversky, "On prediction and judgement," *ORI Research Monograph*, 1972, *12* (4).

_ and _ , "On the psychology of prediction.," *Psychological review*, 1973, *80* (4), 237.

Koehler, Derek J and Greta James, "Probability matching and strategy availability," *Memory & cognition*, 2010, *38* (6), 667–676.

Koehler, Jonathan J, "The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges," *Behavioral and brain sciences*, 1996, *19* (1), 1–17.

Lindeman, Stephan T, Wulfert P Van Den Brink, and Johan Hoogstraten, "Effect of feedback on base-rate utilization," *Perceptual and Motor Skills*, 1988, *67* (2), 343–350.

Manis, Melvin, Ismael Dovalina, Nancy E Avis, and Steven Cardoze, "Base rates can affect individual predictions.," *Journal of Personality and Social Psychology*, 1980, *38* (2), 231.

Medin, Douglas L and Stephen M Edelson, "Problem structure and the use of base-rate information from experience.," *Journal of Experimental Psychology: General*, 1988, *117* (1), 68.

Newell, Ben R and Tim Rakow, "The role of experience in decisions from description," *Psychonomic Bulletin & Review*, 2007, *14* (6), 1133–1139.

_ , Derek J Koehler, Greta James, Tim Rakow, and Don Van Ravenzwaaij, "Probability matching in risky choice: The interplay of feedback and strategy availability," *Memory & Cognition*, 2013, *41* (3), 329–338.

Nisbett, Richard E, Eugene Borgida, Rick Crandall, and Harvey Reed, "Popular induction: Information is not necessarily informative," 1976.

Prasnikar, Vesna and Alvin E Roth, "Considerations of fairness and strategy: Experimental data from sequential games," *The Quarterly Journal of Economics*, 1992, *107* (3), 865–888.

Roth, Alvin E and Ido Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and economic behavior*, 1995, *8* (1), 164–212.

Selten, Reinhard and Rolf Stoecker, "End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach," *Journal of Economic Behavior & Organization*, 1986, *7* (1), 47–70.

Stahl, Dale O, "Rule learning in symmetric normal-form games: theory and evidence," *Games and Economic Behavior*, 2000, *32* (1), 105–138.

Vulkan, Nir, "An economists perspective on probability matching," *Journal of economic surveys*, 2000, *14* (1), 101–118.

West, Richard F and Keith E Stanovich, "Is probability matching smart? Associations between probabilistic choices and cognitive ability," *Memory & Cognition*, 2003, *31* (2), 243–251.

Zukier, Henri and Albert Pepitone, "Social roles and strategies in prediction: Some determinants of the use of base-rate information.," *Journal of Personality and Social Psychology*, 1984, *47* (2), 349.