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

Ignacio Esponda (UCSB)
Emanuel Vespa (UCSB)
Sevgi Yuksel (UCSB)

June 25, 2020

Abstract

We study whether suboptimal behavior can persist in the presence of feedback and examine the role that incorrect mental models play in this persistence. Focusing on a simple updating problem, we document using a laboratory experiment the evolution of beliefs in response to feedback. We compare a baseline treatment, in which a majority of subjects display base-rate neglect (BRN) in initial beliefs, to a control treatment that does not allow for BRN as a mental model but in which learning from feedback is similarly possible. Learning is slow and partial in the baseline, such that after 200 rounds of feedback, beliefs in this treatment are farther from the Bayesian benchmark relative to the control treatment. The treatment effect is linked to partial attentiveness to feedback by those subjects who initially display BRN in the baseline. Presenting subjects with evidence that unequivocally challenges their beliefs by summarizing feedback up to that point improves the accuracy of beliefs substantially and eliminates base-rate neglect. Finally, we find evidence that learning from feedback can generate insights (for example, that the base-rate should be considered in the belief formation process) that can be partially transferred to new settings.

*We thank Jim Andreoni, Ted Bergstrom, Erik Eyster, Guillaume Fréchette, Muriel Niederle, Kirby Nielsen, Ryan Oprea, Collin Raymond, Joel Sobel, Charlie Sprenger, and seminar participants at UCSB, CalPoly, UCSD, UCL, CESS and Stanford for helpful comments. We thank Vincent Guan for research assistance. We are grateful for financial support from the UCSB Academic Senate. We Esponda: iesponda@ucsb.edu. Vespa: vespa@ucsb.edu. Yuksel: sevgi.yuksel@ucsb.edu.

1 Introduction

Behavioral economics has accumulated a wealth of evidence documenting systematic biases in decision making. Some well-known examples include base-rate neglect (Kahneman and Tversky, 1973; Bar-Hillel, 1980), overconfidence (Moore and Healy, 2008; Mobius, Niederle, Niehaus, and Rosenblat, 2011), the sunk-cost effect (Thaler, 1980; Arkes and Blumer, 1985), the law of small numbers (Rabin, 2000), under-exploring (Schotter and Braunstein, 1981; Cox and Oaxaca, 2000), and correlation neglect (Eyster and Weizsäcker, 2010; Enke and Zimmermann, 2019). An important question is whether such biases can persist in the presence of feedback. On the one hand, these biases may vanish with experience if agents are accumulating evidence that is informative of optimal behavior, and adjust their behavior in response to it. On the other hand, convergence to optimal behavior critically presumes people to be fully attentive to the feedback in how they record, process, and incorporate this information to their decisions. A growing empirical and theoretical literature emphasizes how initial misconceptions can have long-lasting effects on how people learn from their experiences.[1] Either due to an incorrect understanding of the value of this information or driven by a desire to hold on to certain types of beliefs, people might not engage with the feedback available to them and thus fail to learn from their experiences. We broadly refer to such failures in incorporating relevant information as resulting from incorrect 'mental models.'[2]

The goal of this paper is to investigate whether suboptimal behavior can persist in the presence of frequent and abundant feedback and, specifically, to assess the role that mental models play in this persistence. We do so by designing a laboratory experiment with two crucial features. First, the decision maker faces 200 rounds of the same decision problem and receives transparent feedback that is informative and simple (natural sampling). This allows us to test for persistence of behavior after significant experience. Second, we consider a baseline treatment that induces an incorrect mental model in many subjects' minds, and compare it to a treatment in which feedback of the same quality is provided but where, by design, the incorrect model is absent. This design allows us to study the extent to which initial misconceptions can play a role in inhibiting learning.

As a proof of concept, we focus on one of the most well-documented biases in the literature, base-rate neglect, to induce an incorrect mental model. Base-rate neglect (BRN)

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

[2] We borrow the term 'mental model' from the psychology literature (see Johnson-Laird (1980) for early use of the term). In economics, the term often refers to misspecified models (most recently used in Graeber (2019)).

describes the tendency to underuse prior beliefs (the base rate) when updating in light of new information relative to the Bayesian benchmark. As a motivating example (adapted from Kahneman and 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.[3] With these primitives, the chance that the person is sick conditional on a positive test result is 41 percent, but the literature has repeatedly documented that many subjects (and doctors!) incorrectly consider this chance to be 80 percent (see Benjamin (2019) for a survey). Because such beliefs *completely* fail to take into account the unconditional probability of the disease, we refer to this bias as *perfect* base-rate neglect (pBRN).

Our experimental design involves many repetitions of the updating problem described in the motivating example above (but presented using a more neutral framing). We first replicate standard findings that most subjects' initial beliefs are consistent with pBRN. Then, subjects face the same decision problem for 200 rounds. In each round, a new state is randomly selected and a signal is drawn. Subjects submit beliefs conditional on the signal, and observe the true state at the end of the round. The interface also displays a record of all past outcomes. In our baseline treatment, subjects are informed of the primitives (i.e., the 15 percent prior and the 80 percent accuracy of the signal) so that, in principle, they could provide the correct response of 41 percent (conditional on a positive signal) from the very first round. Our focus is not on the precise dynamics of how beliefs change in response to feedback, but on whether beliefs by round 200 are close to the Bayesian benchmark. 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 65 percent in round 1, drops to slightly below 55 percent by round 200. While the adjustment is significant, it also remains substantially above the Bayesian benchmark of 41 percent.

To study the degree to which incorrect mental models (BRN in our case) can impact the consistency of beliefs with Bayesianism in the presence of feedback, we need a counterfactual environment where this incorrect mental model is not induced, but where learning from feedback is similarly possible. With this aim, we conduct a control treatment in which subjects face the same updating task described in the baseline, 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

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

generated by the same primitives, and it is exogenous to the subjects' beliefs. However, since they are not provided with the primitives, we do not induce the incorrect mental model (BRN) of the baseline treatment. Hence, this control treatment allows us to observe how subjects' beliefs evolve in the long run in response to feedback in the absence of this mental model. The experimental design we propose to study the effect of mental models on learning is one of the contributions of the paper and can potentially be applied to study the persistence of other biases of interest.

We find an important treatment effect after 200 rounds with respect to the accuracy of beliefs: In aggregate, beliefs in the control treatment (without primitives) are closer to the Bayesian benchmark relative to beliefs in the baseline treatment (with primitives). For example, the average belief conditional on a positive signal is at 45 percent in the control treatment (without primitives) which is 10 percentage points lower than the value in the baseline treatment (with primitives).

We then take a closer look at individual behavior to study how mental models may hinder learning from feedback. Specifically, we focus on a subset of subjects in the baseline treatment whose initial beliefs are consistent with pBRN, i.e., they completely neglect the prior. By round 200, beliefs of these subjects are farther away from the Bayesian benchmark compared to other subjects in the same, baseline treatment. We also identify these subjects to be the main drivers of the aggregate treatment effect relative to the control treatment in which subjects were not told the primitives. While the magnitude of the bias in beliefs is attenuated through feedback, subjects whose initial beliefs are consistent with pBRN still show evidence of base-rate neglect by round 200. Our findings thus indicate that an incorrect mental model (as captured by pBRN) is the main driver behind the persistence of biased beliefs.

We then turn to understanding the channels through which an incorrect mental model (such as base rate neglect) can hinder learning. We document that the subjects in the baseline treatment whose initial beliefs are consistent with pBRN are less responsive to immediate and cumulative feedback relative to others in the same treatment and relative to subjects in the control treatment (without primitives). In addition, they spend less time per choice compared to subjects in the control treatment. We also show that these subjects have a less accurate recollection of the feedback they experienced in 200 rounds. Overall, these patterns suggest these subjects to be less attentive to the feedback relative to others in the baseline as well as those in the control treatment. These findings are consistent with the idea that subjects who develop a mental model early on are less inclined to subsequently examine the data.

Next, we test whether subjects respond differently to feedback when it is presented to them

in a summarized form. Specifically, after being asked to recall the data at the end of round 200, the subjects are provided with a correct summary of all the data they have experienced, presented in an easy-to-read table format. We find that this is sufficient for eliminating the treatment effect. This finding further supports the idea that subjects who develop a mental model early on are inattentive to the data. But, importantly, these subjects are able and willing to modify their mental model when confronted with unequivocal evidence that goes against it.

Finally, we assess the extent to which learning is transferable across environments. In particular, while we find that feedback is instrumental in shaping subjects' beliefs and correcting misperceptions, we ask to what extent subjects learned that their mental model was incorrect because it neglected the information on the prior. Indeed, we find evidence that learning is partially transferable across environments: average beliefs incorporate the information on the prior more in the new environment, but a non-negligible amount of base-rate neglect remains.

Our findings have several implications. The exercise can be thought as a 'proof of concept' example that can be applied to establish to what extent mental models in other settings are resilient to feedback. More broadly, our results provide insights on how incorrect mental models can persist in the presence of abundant feedback. That is, people with an incorrect mental model appear to be less likely to use feedback to inform their choices, and that it might indeed be the presence of the mental model that prevents them from being attentive to the feedback. However, providing subjects with evidence that runs counter to their mental model, even if it is evidence that they have already encountered in scattered form before, can be instrumental in shifting behavior towards optimal choices.

These observations have implications for how policies should be designed to counteract behavioral biases. First, our results suggest that biases can be persistent even in information rich environments where optimal behavior is easy to identify. Hence, to successfully mitigate these biases, in addition to providing agents with information, policy interventions would need to influence how agents engage with this information. An interesting implication of our results is that withholding payoff relevant information (as in the control treatment where subjects are not told the primitives) can lead to long-run choices that are closer to optimal in contexts where such information is likely to induce incorrect mental models. In terms of the specific bias that we study, our findings suggest that more attention should be paid to theories that allow for agents who exhibit partial base-rate neglect (for a recent example, see Benjamin, Bodoh-Creed, and Rabin, 2019).

Throughout the paper, we use the term 'mental model' broadly to refer to an agent's possibly incorrect initial understanding of the environment. Incorrect mental models may not

only generate suboptimal behavior in the short run, but they can also impact behavior in the long run by influencing how attentive the agent is to information about past experiences. We are also using 'attentiveness' in a general way to encompass any frictions in how the agent acquires, processes and records information about past outcomes. In this sense, the themes explored in this paper, in terms of how learning from past experiences is necessarily shaped by our initial understanding of the world, connect with a few different literatures as we outline below.

First, our results provide support for a growing literature in economics that studies the implications of incorrect, misspecified mental models. A central premise of this literature is that the degree to which an agent learns from past experiences is constrained by her initial misspecified model.[4] Some of this work connects such representations to models of behavioral agents as developed by Gennaioli and Shleifer (2010), Bordalo, Gennaioli, and Shleifer (2013), and Gabaix (2014). A related literature also emphasizes cognitive difficulties associated with comprehending and integrating important features of the environment to the decision making process.[5] Such cognitive difficulties may explain agents' reliance on simpler (but incorrect) mental models.

Second, an emerging literature endogenizes attentiveness to payoff-relevant features of the environment when there are information processing costs. The literature on rational inattention (e.g., Sims, 2003; Caplin and Dean, 2015) assume agents have rational expectations about the value of such information, but trade off this value against learning costs. Building on this intuition, but allowing agents to be systematically misguided in how they assess the value of information (in the tradition of misspecified models), Schwartzstein (2014) and more recently Gagnon-Bartsch, Rabin, and Schwartzstein (2018) 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 incorrect mental models 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.[6]

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

[5]See for example, Eyster and Weizsäcker (2010), Cason and Plott (2014), Esponda and Vespa (2014), Louis (2015), Dal Bó et al. (2018), Ngangoue and Weizsäcker (2018), Esponda and Vespa (2019), Martínez-Marquina, Niederle, and Vespa (2019), Araujo et al. (2019), Martin and Muñoz-Rodriguez (2019), Moser (2019), Graeber (2019), Enke and Zimmermann (2019), Enke (2019), Bayona, Brandts, and Vives (2020).

[6]For example, subjects in our baseline treatment who adopt the pBRN mental model may be less willing to pay a mental cost to process the feedback because they incorrectly believe that this data cannot improve the optimality of their answer. Following the language of Handel and Schwartzstein (2018), such failures in learning

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 and Tirole, 2003; Brunnermeier and Parker, 2005; Köszegi, 2006) or simply due to a desire for consistency (Falk and Zimmermann, 2018), agents might be reluctant to adjust their behavior in response to past outcomes.[7] Note that these different literatures share a common insight that initial misconceptions can inhibit learning by impacting the way agents engage with the data. While our experiment provides strong evidence for this common channel, it is not designed to distinguish between the different ways of formalizing mental models in the literature, and we believe this task is better left for future work.

Our paper also relates to a literature that studies long-run outcomes in the presence of feedback, often in environments where well-known biases play a role. While we would expect incorrect mental models to impact learning in many of these settings, properly identifying this channel can be challenging. For example, learning in strategic settings is complicated by the fact that agents may also have to make inferences about the strategies of others, and these strategies may change over the course of the experiment. Moreover, in many problems, feedback is often partial, noisy (e.g. Huck, Jehiel, and Rutter, 2011), or more importantly, endogenous to the subject's choices. Learning can also be cognitively challenging if agents face a dataset with sample selection (e.g. Esponda and Vespa, 2018; Araujo, Wang, and Wilson, 2019; Barron, Huck, and Jehiel, 2019). Yet another example of why learning from feedback might be difficult is the case of an agent who (given her model of the world) makes choices such that the collected information does not challenge her understanding of the world (e.g. Dekel, Fudenberg, and Levine, 2004; Fudenberg and Vespa, 2019). To control for these issues, we focus on a decision problem in which feedback is simple, transparent and exogenous to the subjects' choices.

There is also a large literature on the specific bias that we study, base-rate neglect, initiated by Kahneman and Tversky (1972); this literature has been recently surveyed in Benjamin (2019). 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 lit-

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. There is growing empirical evidence 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, and Schwartzstein (2018) for more examples.

[7] See Bénabou and Tirole (2016) for an extensive review of this literature. Recently, Zimmermann (2018) and Huffman et al. (2018) study the connection between persistent overconfidence and distortions in memory through selective recall when there is repeated feedback.

erature focuses on the evolution of beliefs when subjects are not given the primitives and only observe outcomes from a natural sampling process. The paradigm in this literature is to study the *description-experience gap* which compares accuracy of beliefs when subjects are only given the primitives to when subjects only have experience to rely on. For example, Gigerenzer and Hoffrage (1995) show that base-rate neglect is attenuated when subjects are provided with natural frequencies (instead of the underlying primitives). To our knowledge, there has not been an experiment contrasting learning in treatments with and without primitives with the goal of studying the role mental models play in the persistence of biases.[8]

2 Experimental Design

2.1 Procedures and Treatments

I. Updating task: Round 1

The experiment consists of five main parts.[9] This first part, referred to as round 1, introduces the main belief-updating task. The task consists of updating beliefs on a binary state using a binary signal. Our experimental design consists of two between-subject treatments which differ only in the instructions provided in this part. The treatments, referred to as *Primitives* and *NoPrimitives*, vary in whether subjects are provided with the primitives of the problem or not. Subjects are told in both treatments that there are 100 projects, each either a success or a failure, and the task consists of assessing the chance that a randomly selected project is a success vs. a failure conditional on a signal that is informative about the type of the project. In *Primitives*, subjects know that 15 projects are successes and 85 projects are failures. In *NoPrimitives*, subjects know that some projects are successes and some are failures, but they are *not* told how many are successes and how many are failures. We frame the signal as the computer running a test on the selected project. The signal is either positive or negative. In *Primitives*, subjects also know that the signal has a reliability of 80 percent.[10] In *NoPrimitives*, subjects are told that the signal has a reliability of q percent, but while we describe the meaning

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

[9] For expositional purposes we describe our experiment here in five parts, though the presentation for subjects was broken up into nine parts. See the Online Procedures Appendix for details.

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

of q just as in *Primitives*, we do not reveal the value of q. This parameterization (prior = .15, reliability of signal = .8) is the same for both treatments and corresponds to the classic parameterization of Kahneman and Tversky (1972).

To summarize, the *only* difference between the two treatments is that subjects know the prior and the reliability of the signal in *Primitives*, while these values are not provided to the subjects in *NoPrimitives*. All other parts of the instructions, in this part and in all subsequent parts, are identical. In both treatments, using the strategy method, we ask subjects to submit two assessments: (1) the belief that the project is a success vs. failure conditional on the test being positive (B_{Pos}), and (2) the belief that the project is a success vs. failure conditional on the test being negative (B_{Neg}). In this round and in all future belief-elicitation rounds, subjects are incentivized using a standard incentive-compatible mechanism.[11]

In *Primitives*, subjects could in principle use Bayes' rule to provide the correct answer. But, as the literature has documented, most subjects are not able to provide the correct answer and instead suffer from base-rate neglect (Benjamin et al., 2019). Thus, by providing the primitives in round 1, we induce an incorrect mental model (for some subjects) in *Primitives* which involves neglecting the base rate. In *NoPrimitives*, there is no correct way to respond and there is of course no way to suffer from base-rate neglect, 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 1 for another 199 rounds. This part is divided into two phases. The first phase encompasses rounds 2 through 100. At the end of each round, subjects receive feedback on the signal (test result is positive vs. negative) and state (project is a success vs. failure) realizations. The right side of the screen includes a history box that records the signal and state realizations observed in each of the past rounds. Figure 1 shows a screen shot of round 5. In the top-left of the screen, the subject submits a belief conditional on a positive signal and a belief conditional on a negative

[11] Belief elicitation has been combined with the strategy method in a number of prior information-response experiments, e.g. Cipriani and Guarino (2009), Toussaert (2017), Agranov, Dasgupta, and Schotter (2018), Charness, Oprea, and Yuksel (2019). See Danz, Vesterlund, and Wilson (2020) for a recent evaluation of belief elicitation practices and Online Procedures Appendix B for further details on how our design introduces the elicitation method.

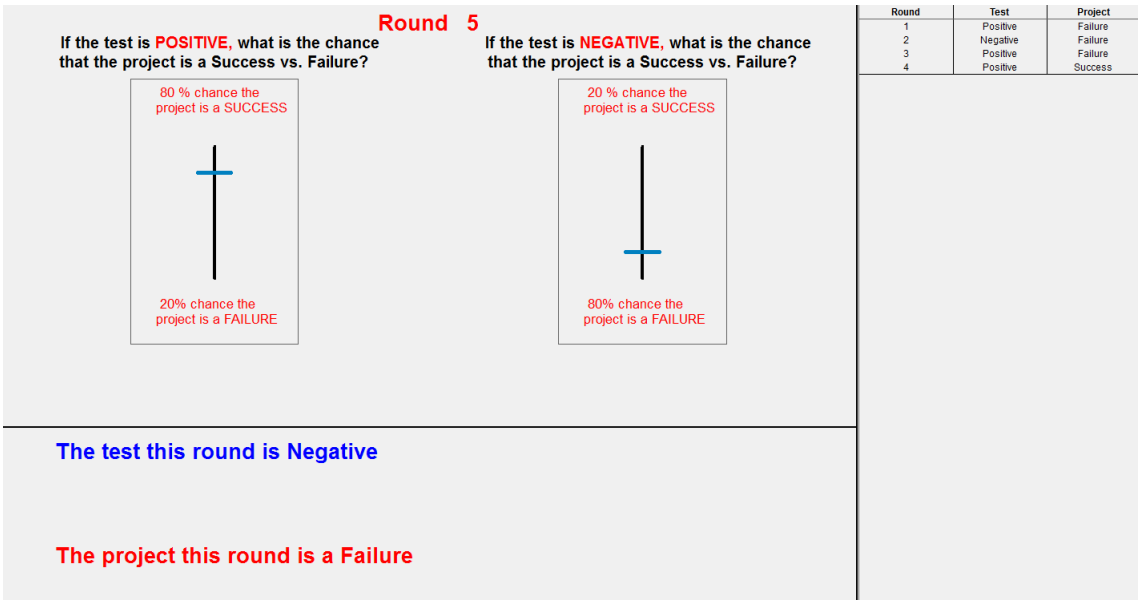


Figure 1: Interface screenshot at round 5

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

The second phase encompasses rounds 101 through 200. The only difference with respect to the first phase is that subjects are asked to report their beliefs only every 10 rounds, as opposed to in every round, while receiving feedback in real time in every round. For example, a subject who just submitted responses for round 110 would see the outcomes for each of rounds 110 through 119 without being asked again for her beliefs until round 120; the same procedure follows in blocks of ten rounds. This is done to be able to assess how an additional 100 rounds of feedback would affect beliefs while keeping the experiment to a reasonable time limit.

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

III. Recollection of feedback

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.

IV. Summary tables

In this part, we confront subjects with the actual data they observed in a conveniently aggregated manner.[12] 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 phases. In the first phase, 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.

In the next phase, 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.

In the final phase, the interface computes the relevant frequencies of the entries presented in the table from the previous phase. 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 do one more round of the belief elicitation task.

Clearly, as the phases increase, subjects with incorrect beliefs are confronted with stronger evidence against their beliefs.

[12] Recall that in previous parts subjects observe the entire past history, so they can see the history of signal-state realizations at all times, but throughout previous parts the data is not summarized for them in the manner we do in this part.

V. Transfer of learning

So far our design does not distinguish between two different ways in which subjects 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 mental model is incorrect. In turn, this type of deeper learning can help improve decision making in related but different environments (where subjects cannot rely on previously accumulated data). In the last part of the experiment, we study whether subjects engage in this type of deeper learning. In particular, we assess the extent to which subjects learned that their mental model was incorrect because it neglected the information on the prior. To do so, 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.

Experimental procedures

As mentioned earlier, we conducted a between-subjects design with two treatments, *Primitives* and *NoPrimitives*. As the names indicate, the only difference between these treatments is that, in the former, subjects are informed of the prior and the reliability of the signal for all parts of the experiment. Subjects encounter each part of the experiment in sequence and do not know what parts they will encounter in the future. For example, when starting round 2, they only know that they will be making this type of decision for rounds 2 through 100, but they are not told that there will later be 100 more rounds. Subjects are allowed to make choices at their own pace. To prevent rushed decisions as much as possible, we informed subjects (and enforced the rule) that if they wanted to be paid (in addition to their show up payment) they would not be allowed to leave the laboratory before the 90 minute mark.

Before subjects began round 1, we introduced them to the belief elicitation task and the incentive-compatible mechanism using simple examples. For each subsequent task, we also provided subjects with detailed instructions and then tested their comprehension with questions. 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).

We provide more details about the experimental procedures in Online Appendix B. For the

full details that allow an exact replication of our experiment, we refer the reader to the Online Procedures Appendix, where we include instructions and screenshots relating to each part.

The experiments were conducted at the University of California, Santa Barbara and subjects were recruited using ORSEE (Greiner, 2015). In total, 128 subjects participated in the experiment (64 in each treatment). The experiment, which lasted 90 minutes, was conducted using zTree (Fischbacher, 2007). In addition to the \$10 show up payment, earnings from the experiment were either \$25 or \$0, for a grand total of either \$10 or \$35.[13] Payments on average equaled \$22.5.

2.2 Bayesian and Base-rate neglect benchmarks

Given the prior $p = .15$ (ex-ante probability that the project is a success) and the reliability of the signal, $q = .8$, the Bayesian posterior that the project is a success conditional on a positive signal is, in percentage terms, $B^*_{Pos} = \frac{pq}{pq+(1-p)(1-q)} \times 100\% = 41\%$. Similarly, the Bayesian posterior that the project is a success conditional on a negative signal is $B^*_{Neg} = 4\%$. Let (B_{Neg}, B_{Pos}) capture the subject's reported beliefs in percentage terms, namely, their assessment that the project is a success conditional on a negative and positive signal, respectively. Throughout the paper, we refer to a subject's beliefs as Bayesian (in a given round/part) if $(B_{Neg}, B_{Pos}) = (B^*_{Neg}, B^*_{Pos})$.[14]

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

In the last part of the experiment, the new primitives are $p' = .95$ and $q' = .85$, and the corresponding Bayesian and pBRN beliefs are $(B^{*'}_{Neg}, B^{*'}_{Pos}) = (77, 99)$ and $(B^{pBRN'}_{Neg}, B^{pBRN'}_{Pos}) = (15, 85)$.

[13] 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.

[14] Note that subjects in *Primitives* could, in principle, submit Bayesian posteriors in every round of the experiment where they are faced with the main updating task. Since the primitives (values of p and q) are not provided to the subjects in *NoPrimitives*, this was clearly not possible in every round, but only approximately possible in the long run.

2.3 Understanding the design

We designed the experiment to serve several goals. First, the design allows us to study the persistence of a well-documented bias (BRN) in the presence of feedback in a simple framework. Responses in round 1, where the main task is first introduced, provide a benchmark for beliefs in the absence of feedback. Evolution of beliefs in later parts allow us to observe the impact of feedback. Feedback is natural (the state-signal realization of that round), informative and independent of the subjects' choices.

Second, the design includes a control treatment (without primitives) in which feedback is structurally the same, but the dominant incorrect mental model of the baseline treatment (BRN) is not present. Thus, the control treatment provides us with a benchmark on how subjects' beliefs would evolve in the long run in the absence of this mental model, but where subjects can still rely on feedback. The experimental design we propose in order to study the effect of mental models on learning is one of the contributions of the paper and can potentially be applied to study the persistence of other biases of interest in the future.

Third, we added several design features to study whether mental models impact the way subjects engage with the feedback: (i) we observe response to immediate and cumulative feedback over the course of 200 rounds; (ii) we keep track of response times; (iii) we ask subjects to recall the feedback they experienced; and (iv) we test whether subjects' response to feedback changes when it is presented to them in a summary form that unequivocally challenges the BRN model.

Finally, we can use the last part of the experiment to assess the extent to which learning in one environment is transferable to other environments.

3 Results

We organize our main results as follows: In Section 3.1, we confirm that initial (i.e., round 1) responses in *Primitives* replicate previous findings in the literature related to BRN. In Section 3.2, focusing on 200 rounds of feedback, we document differences between *Primitives* and *NoPrimitives* at the aggregate level. We show that, by round 200, beliefs in *NoPrimitives* are closer to the Bayesian benchmark than beliefs in *Primitives*. In Section 3.3, we show that these treatment effects are largely driven by those subjects in *Primitives* whose initial beliefs are consistent with pBRN, suggesting that incorrect mental models play an important

role in slowing down convergence towards the Bayesian benchmark. Sections 3.4 and 3.5 provide evidence that these subjects are less attentive to the feedback, but that presenting feedback in a summarized table form, which clearly challenges the pBRN model, is sufficient for eliminating the treatment effect. Finally, in Section 3.6, we study the degree to which learning in one environment with one set of primitives transfers over to another with new primitives.

3.1 Replication in treatment *Primitives*

Here, we summarize patterns in initial responses in *Primitives*, that is, round 1 of the updating task. The mode and the median belief reported conditional on a positive signal (B_{Pos}) is 80 percent (the pBRN prediction), which is consistent with the results for the same parameterization in Kahneman and Tversky (1972).[15] 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 possible computation errors.[16] Besides the pBRN and Bayesian benchmarks, another natural response involves signal-neglect, where beliefs conditional on either signal coincide with the prior. We find that 7.8 percent of our subjects respond in this way.

In the upcoming sections, we will present more details on the distribution of beliefs in *Primitives* and contrast it to *NoPrimitives*. The main message from this section is that the baseline condition needed for our study holds: For most subjects in *Primitives*, beliefs submitted in the first round are consistent with pBRN. We interpret this as *Primitives* inducing an incorrect mental model for many subjects. Next, we study choices in rounds 1-200 at the aggregate level to evaluate to what extent feedback can correct such behavior.

[15] Kahneman and Tversky (1972) only ask about beliefs conditional on the signal for which the pBRN response is 80 percent.

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

3.2 Results at the aggregate level

A first look at the results

We start by describing aggregate-level behavior for different rounds in *Primitives* to evaluate if providing natural feedback can eliminate the bias.[17,18] Figure 2a presents the evolution of beliefs in *Primitives* using red squares–where the number next to a square indicates the round that the average corresponds to. The average round 1 beliefs in *Primitives* are $(B_{Neg}, B_{Pos}) = (22, 64)$. We observe that average beliefs in this treatment move closer to the Bayesian benchmark (and away from the pBRN point) with experience: After 100 rounds, average beliefs are $(16, 53)$, which corresponds to an adjustment towards the Bayesian benchmark of about six and eleven percentage points in B_{Neg} and B_{Pos}, respectively. However, at this point, average beliefs are still twelve percentage points away from the Bayesian benchmark conditional on either signal.

The evidence suggests that while beliefs move towards the Bayesian benchmark with experience, the adjustment is slow and partial after 100 rounds. Note, however that there could be many factors that slow down learning in such a setting. The *NoPrimitives* treatment serves as a natural benchmark allowing us to contextualize results from *Primitives*.

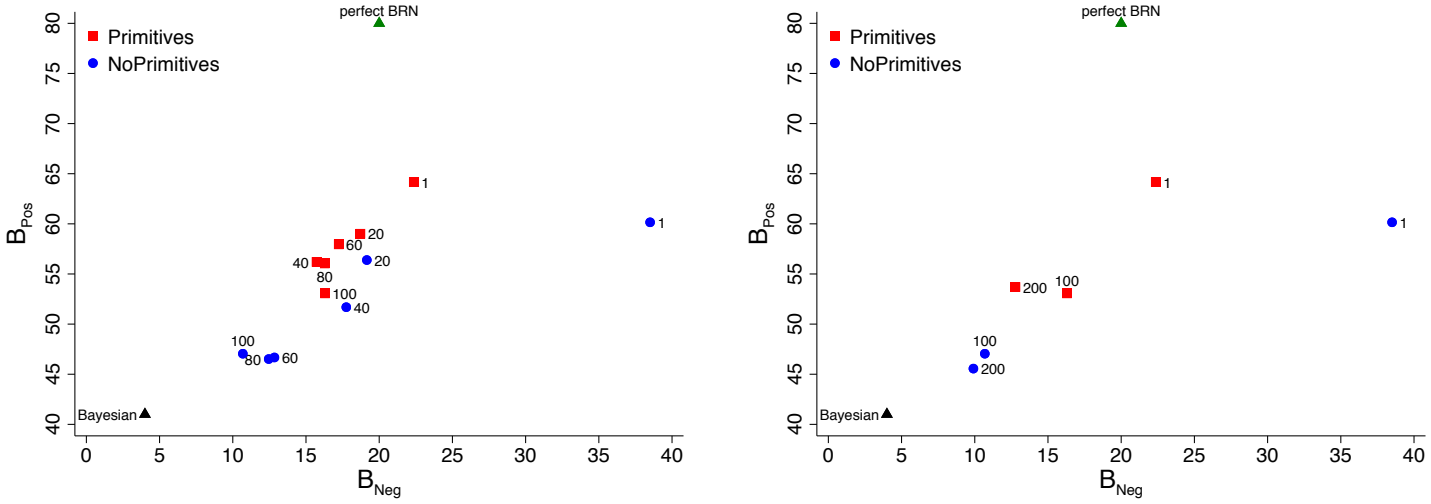
Figure 2a presents average beliefs in *NoPrimitives* for different rounds using blue circles. Average beliefs in round 1 equal $(39, 60)$, which is relatively far from the Bayesian benchmark. Yet after 100 rounds beliefs move close to the benchmark, reaching $(11, 47)$. Figure 2a indicates that after 100 rounds there is a treatment effect of approximately six percentage points in each dimension. That is, a first look at the evidence suggests that learning is slower in *Primitives* and that having access to the primitives can hinder learning. The figure reveals a similar conclusion if we look at rounds 101-200.

Treatment effects at the aggregate level

In this subsection, we provide a statistical analysis and show that the main comparisons between treatments illustrated in Figure 2 are statistically significant. The reader not interested in the details can skip to the next subsection. Our analysis focuses on providing answers to two key questions: (1) Are there treatment differences in terms of how far beliefs are from the

[17] 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).

[18] Throughout the paper we report beliefs in percentage points.



(a) Round 1 to 100

(b) Rounds 1, 100 and 200

Figure 2: Evolution of beliefs

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

Bayesian benchmark? (2) Are beliefs overall different between the two treatments?[19] We will adopt the following approach to evaluate statistical significance with regards to these questions. To answer the first question, we use a distance measure. Specifically, for a given feedback level and for each subject we calculate $\Delta = \frac{\left|B_{Neg}-B^*_{Neg}\right|-\left|B_{Pos}-B^*_{Pos}\right|}{2}$, which captures the average absolute distance between the subject's submitted beliefs (B_{Neg}, B_{Pos}) and the Bayesian benchmark $(B^*_{Neg,}B^*_{Pos})$.[20] We then estimate the following distance-to-benchmark regression at different feedback levels: $\Delta = a + bP + \epsilon$, where P is a treatment dummy that takes value 1 if the observation is from *Primitives*, and ϵ is a noise term. The coefficient a captures the average distance between beliefs in *NoPrimitives* and the Bayesian benchmark, while b measures the treatment effect, that is, the differential distance between beliefs in *Primitives* and the benchmark relative to *NoPrimitives*.

[19]Note that (1) and (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 directions).

[20]In OnlineAppendix C, we conduct a number of robustness tests for our findings and all results are qualitatively in line with our reports in the main text. First, we also measure distance relative to realized frequencies. Specifically, let F_{Pos} (F_{Neg}) be the observed frequency of success conditional on the signal being positive (negative). The alternative measure of distance uses F_{Pos} and F_{Neg} instead of the Bayesian benchmark. Results using this alternative distance measure are reported in Table 7. Second, we also use Euclidean distance as a measure of distance and we report findings in Table 8. Finally, we include our survey questions as controls and report this in Table 9.

Sample	(1) Dep. var: Δ (distance to benchmark)				(2) Dep. vars: B_{Neg}, B_{Pos}				
	a		b		γ_{Neg}		γ_{Pos}		H_0:
	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	$\gamma_{Neg} = \gamma_{Pos} = 0$
Round 1	27.0	.000	-1.7	.203	-16.1	.000	4.0	.258	.000
Round 100	13.0	.000	5.6	.003	5.6	.035	6.0	.159	.056
Round 200	10.5	.000	4.8	.004	2.9	.112	8.1	.021	.049

Table 1: Estimation output

Notes: Each row presents the results for two sets of regressions. Columns under (1) report the estimates (coeff. column) and p-values (for the null that the corresponding estimate is zero) of: $\Delta = a + bP + \epsilon$, where ϵ is a noise parameter and P a treatment dummy (1 if the observation belongs to *Primitives*). Columns under (2) report the γ estimates and corresponding p-values for: $B_{Neg} = \delta_{Neg} + \gamma_{Neg}P + \upsilon_{Neg}$ and $B_{Pos} = \delta_{Pos} + \gamma_{Pos}P + \upsilon_{Pos}$, which are estimated jointly using the seemingly unrelated regressions procedure. The last column reports a Wald test in which the null hypothesis is that $\gamma_{Neg} = \gamma_{Pos} = 0$. Each row constrains the sample to the beliefs referred to in the first column. Each regression involves 128 observations (64 from each treatment).

To answer the second question, we will use a pair of regressions with beliefs conditional on each signal being the dependent variable in each regression. Specifically, at a given round, for each possible signal $j \in \{Neg, Pos\}$, we run the following regression: $B_j = \delta_j + \gamma_j P + \upsilon_j$, where υ_j is an error term. Because each subject submits a pair of beliefs $(B_{Neg,}B_{Pos})$, we estimate the regressions jointly using seemingly unrelated regressions.[21] Estimating the regressions as a system of equations allows us to test the null hypothesis that both treatment coefficient estimates are equal to zero (i.e. $\gamma_{Neg} = \gamma_{Pos} = 0$).

The results of the distance-to-benchmark and belief regressions at different feedback levels are presented in Table 1. In round 1, beliefs on average in both *Primitives* and *NoPrimitives* are different from the Bayesian benchmark (the estimate for a is at 27 percentage points). In round 1 we can also reject the null that both treatment coefficients from the belief regressions are equal to zero (as can be seen in the last column of Table 1 for R1), indicating that on average beliefs are significantly different between *Primitives* and *NoPrimitives*.[22]

Looking at later rounds, we observe two important patterns. First, as subjects experience more feedback, beliefs in both treatments move closer to the Bayesian benchmark: The estimate for a slowly changes from 27.0 (in round 1) to 13.0 (in round 100) and then to 10.5 (in round 200). Still, by round 200, beliefs in both treatments are significantly far from the Bayesian benchmark.

The second observation is with respect to the treatment differences. By round 200, the

[21] With this procedure, errors can be correlated across equations for a fixed subject, but errors are independent across subjects.

[22] While there is a difference in round 1, by round 20 there is no longer a treatment effect. For details see Table 6 in Online Appendix.

treatment effect in beliefs is close to 8 percentage points in B_{Pos} and 3 percentage points in B_{Neg}. Testing jointly, by round 200 we can reject the null that both treatment effects on beliefs are equal to zero (p-value .049). Focusing on the distance to the Bayesian benchmark, by round 100, beliefs in *NoPrimitives* are closer to the Bayesian benchmark than beliefs in *Primitives*. This pattern also holds true in round 200 (p-value of .003 and .004 for rounds 100 and 200, respectively). Our data indicates that subjects who are not given the primitives to the updating problem end up with beliefs which are on average closer to the Bayesian benchmark than those subjects who are given the primitives!

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 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.[23]

While there are no significant differences in the estimates of β between treatments (and estimates are relatively close to 1), Figure 3 reveals large differences in the estimates of α.[24]

[23]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.

[24]The estimates for β are presented in Figure 10 of Online Appendix C.

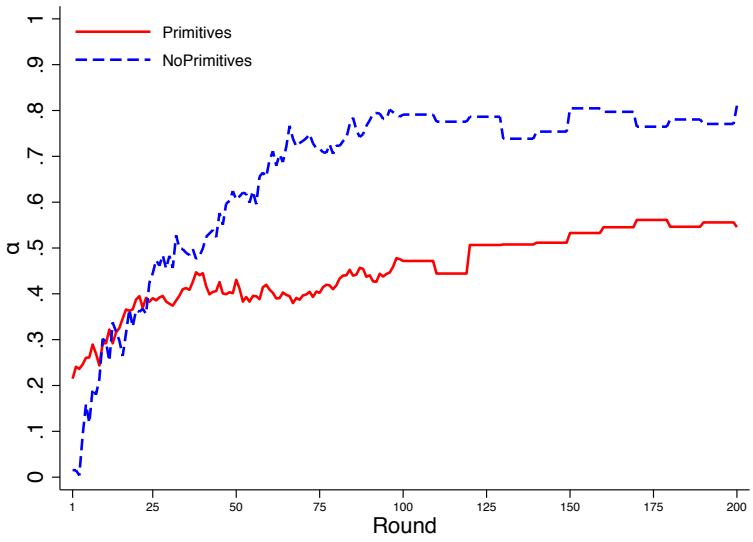


Figure 3: Estimates of α per round by treatment

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 .54, is significantly smaller than that of treatment *NoPrimitives*, which is .82 (p-value $<$.001).

Summary

We next summarize our results on the impact of 200 rounds of feedback:

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

3.3 The role of an incorrect mental model in hindering learning

The aggregate treatment effects documented above are consistent with our hypothesis that incorrect mental models can hinder learning from feedback. However, it is possible that those subjects who know the primitives are unable to learn from feedback for other reasons that are independent of any mental model. To further assess the role of mental models, we take a

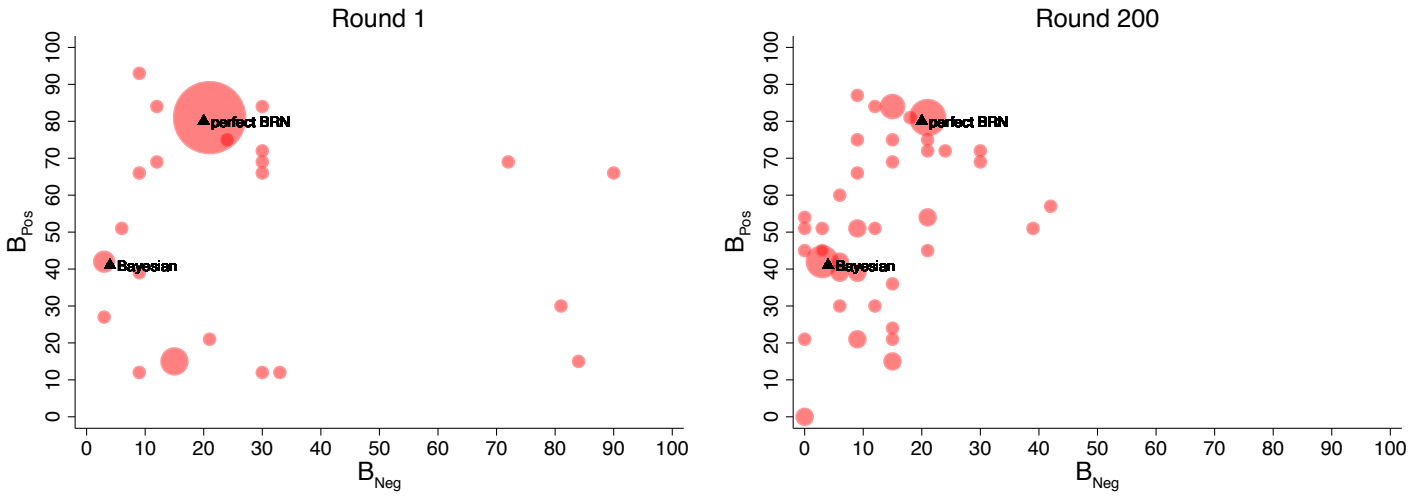


Figure 4: Density plots for *Primitives*

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.

closer look at the data to account for heterogeneity. If the mental model (pBRN) were responsible for hindering learning, we should see that the aggregate treatment effects documented in the previous section are driven by those subjects in *Primitives* whose initial responses are consistent with pBRN. This is exactly what we document in this section.

As a precursor to our main finding in this section, we begin by providing an overview of the heterogeneity in responses. Figures 4 and 5 present the distribution of beliefs in *Primitives* and *NoPrimitives* at different feedback levels. As can be seen in the left plot of Figure 4, the majority of subjects (56.3 percent) in round 1 of treatment *Primitives* submit beliefs that are consistent with pBRN. There are a few other points around which beliefs are somewhat concentrated. For instance, 4.7 percent of subjects have Bayesian beliefs, and 7.8 percent of subject display signal neglect (i.e. beliefs conditional on either signal are equal to the prior). The right plot of Figure 4 shows that by round 200 the distribution of beliefs has shifted significantly and that most subjects can essentially be categorized into two groups. There is a large cluster close to or at the pBRN point and another close to or at the Bayesian point.[25]

As can be seen in the left plot of Figure 5, subjects' beliefs in round 1 of *NoPrimitives* are quite different from those of *Primitives*, though they are approximately equidistant to the Bayesian benchmark. These beliefs can largely be organized into two groups. A large mass of

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

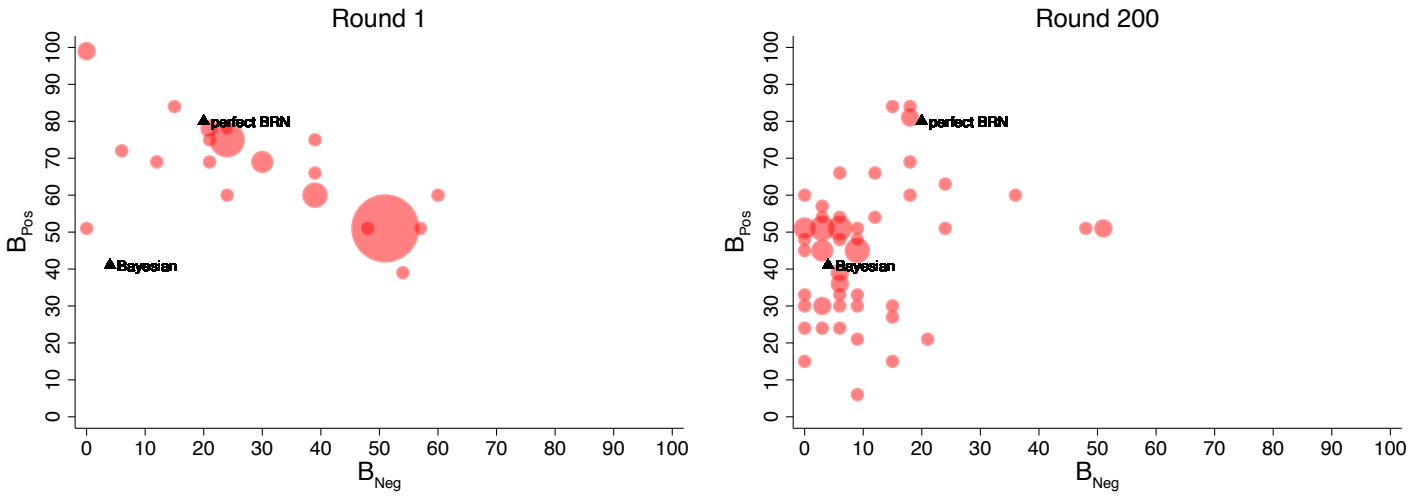


Figure 5: Density plots for *NoPrimitives*

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

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. That is, their beliefs indicate that a randomly selected project is more likely to be a success conditional on a positive vs. a negative signal ($B_{Pos} > B_{Neg}$). By round 200 (right plot of Figure 5), the mass at $(50, 50)$ largely disappears and the vast majority of subjects concentrate around the Bayesian point.[26] We summarize these observations:

Observation: *In Primitives, beliefs for the majority of the subjects are consistent with pBRN in round 1, but are split into two groups by round 200: those close to the pBRN benchmark and those close to the Bayesian benchmark. In NoPrimitives, subjects' beliefs in round 1 are either* $(50, 50)$ *or ranked to reflect informativeness of labels* ($B_{Pos} > B_{Neg}$)*, and by round 200, beliefs for most subjects are close to the Bayesian benchmark.*

While these patterns are suggestive of the (partial) persistence of the pBRN model, the distributions displayed in Figures 4 and 5 do not connect the behavior of individual subjects across rounds. Thus, to evaluate the stability of beliefs across rounds directly, we study separately the evolution of responses for those subjects in *Primitives*, whose beliefs are consistent

[26] Fifty-two percent of subjects are at ± 10 percentage points of the realized frequencies.

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 2: 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.

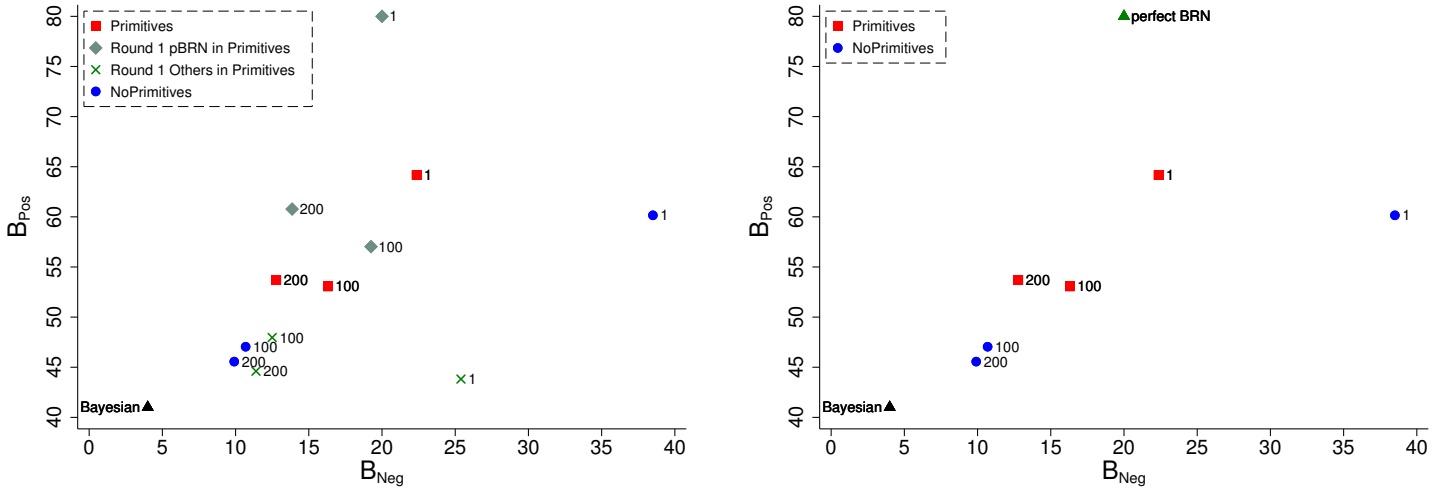
with pBRN in round 1. We will refer to these subjects as *Round 1 pBRN subjects*.

Figure 6a reproduces the same information presented previously in Figure 2b, but in addition also 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 1, so that the starting point for this group is $(B_{Neg}, B_{Pos}) = (20, 80)$. 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 2; for example, there is a significant fifteen percentage-point difference in the average of B_{Pos} between the two groups.[27,28]

In Figure 6b, we demonstrate that these distinct patterns observed for Round 1 pBRN subjects are not due to the fact that they start out in round 1 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

[27] If we test the joint hypothesis that there are differences in B_{Pos} and B_{Neg}, we obtain p-values of .007 and .001 in rounds 100 and 200, respectively.

[28] In this section, 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 13 of Online Appendix C.



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

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

Figure 6: 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.

sample in both treatments to include only subjects with $B_{Pos} \in [70, 100]$ and $B_{Neg} \in [0, 30]$ in round 1. In *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.[29]

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 6a. 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*.[30]

In summary, the decomposition of subjects in *Primitives* depending on their round 1 choices shows that beliefs of Round 1 pBRN subjects in round 200 are statistically differ-

[29]Table 11 in Online Appendix C verifies these patterns statistically.

[30]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 2 equals .011, but the same test for estimatates in column (4) delivers a p-value of .760.

ent 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 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*.

We summarize the main takeaways from this section:

Finding #2: *The aggregate differences between Primitives and NoPrimitives in round 200 are driven by those subjects in Primitives who perfectly neglect the base rate in round 1. All other subjects in Primitives have beliefs that are similar to the beliefs of subjects in NoPrimitives. This finding corroborates that the presence of an incorrect mental model (pBRN) hinders learning from feedback.*

3.4 Attentiveness and treatment effects

So far, we have established that an incorrect metal model (pBRN) hinders learning from feedback, but a question arises as to why the subjects who start out with this incorrect mental model are not able to learn from feedback as efficiently as those who do not start out with such a model. We study this issue by examining how subjects respond to feedback. We provide evidence that subjects in *Primitives*—particularly the Round 1 pBRN subjects who are driving the aggregate treatment effects—are less responsive to immediate feedback or to cumulative feedback, and that their responses are more likely to show convergence (in the sense of beliefs not changing).

Response to immediate feedback

We study how beliefs change in the first 100 rounds, when subjects can make adjustments after each round of feedback.[31] We focus on two regressions, one for each reported belief in which the dependent variable is the difference between the beliefs in round t and $t-1$. On the right-hand, side we include four dummies, each representing whether one of the four possible signal-state realizations occurred in round $t-1$.

Results presented in Table 3 suggest that on average subjects in *NoPrimitives* are highly responsive to immediate feedback. For example, experiencing a signal-state realization of

[31] As can be seen from Figure 2b and Table 1 the changes in average beliefs indeed mostly take place within the first 100 rounds, but patterns presented in this section are robust to looking at all 200 rounds.

	(1) Dep. Var.: $B^t_{Pos} - B^{t-1}_{Pos}$				(2) Dep. Var.: $B^t_{Neg} - B^{t-1}_{Neg}$			
	Primitives			*NoPrimitives*	*Primitives*			*NoPrimitives*
	All	Round 1 pBRN	Round 1 Others	All	All	Round 1 pBRN	Round 1 Others	All
(Pos, Succ)$_{t-1}$	0.7	0.5	1.0	2.4	0.1	0.2	-0.1	0.0
	(.028)	(.201)	(.076)	(.001)	(.770)	(.594)	(.377)	(.894)
(Pos, Fail)$_{t-1}$	-1.1	-0.9	-1.4	-2.8	0.1	0.1	0.1	0.4
	(.008)	(.147)	(.006)	(.000)	(.832)	(.922)	(.344)	(.072)
(Neg, Succ)$_{t-1}$	-0.2	-0.6	0.2	-0.9	0.9	0.6	1.3	2.0
	(.611)	(.476)	(.725)	(.069)	(.007)	(.149)	(.020)	(.004)
(Neg, Fail)$_{t-1}$	0.0	-0.2	0.2	0.1	-0.2	-0.1	-0.3	-0.6
	(.951)	(.275)	(.023)	(.224)	(.045)	(.510)	(.004)	(.000)

Table 3: Changes in reported beliefs after feedback

Notes: Each column in group (1) and (2) provides estimates for regressions in which the dependent variable is the change in beliefs conditional on each signal (B_{Pos} in (1) and B_{Neg} in (2)) from round $t-1$ to t. The right-hand side consists of a set of dummies that covers all possible outcomes in the previous round. For example, (Pos, Succ)$_{t-1}$ is a dummy that takes value 1 if the feedback in the previous period was that the signal was positive and the state (the type of the project) was a success. The first three columns report results for subjects in *Primitives*: first including all subjects ('All'), then including only subjects who were classified as Round 1 pBRN and lastly including subjects who were not classified in Round 1 as pBRN ('Round 1 Others'). The fourth column reports results for all subjects in *NoPrimitives*. Each regression includes data from all beliefs submitted in in rounds 2 to 100. Between parentheses, we report p-values for the null hypothesis in which the coefficient equals zero. Standard errors used in the computation of the p-values are estimated by clustering at the subject level.

positive-success moves B_{Pos} upwards by 2.4 percentage points in the next round. Meanwhile, experiencing a signal-state realization of positive-failure moves B_{Pos} downwards by 2.8 points. A similar pattern appears if we study the adjustment of B_{Neg} for subjects in *NoPrimitives*. In this case, subjects are largely unresponsive if the signal is positive, but adjust their beliefs upwards after a signal-state realization of negative-success and downwards after one that is negative-failure, with both estimates being significant at the one-percent level.

While we find qualitatively similar patterns in *Primitives*, the coefficients are much smaller than those estimated for *NoPrimitives*, particularly when we focus on B_{Pos}. Separating subjects in *Primitives* as Round 1 pBRN and Round 1 Others provides further insight. For Round 1 Others, the qualitative pattern of responses is more similar to those subjects in *NoPrimitives*. By contrast, Round 1 pBRN subjects do not appear to systematically respond to immediate feedback. None of the estimates are significant for this subgroup of subjects and most estimates are relatively small.

Responses to cumulative feedback

While results on responses to immediate feedback reveal differences between Round 1 pBRN subjects in *Primitives* and Others, the analysis does not capture responses to cumulative feedback. It is possible, for example, that Round 1 pBRN subjects do not immediately respond to feedback but adjust their beliefs only after they collected a large number of observations.

To gain some insight into how cumulative feedback influences beliefs, we look at how far reported beliefs are from realized frequencies. Specifically, we use a distance-to-benchmark regression where the dependent variable is the average absolute distance between the subject's beliefs in round 100 and the relevant frequencies observed in feedback up to that point.[32]. The analysis shows that beliefs in *Primitives* are on average 10.8 percentage points further away from realized frequencies relative to beliefs in *NoPrimitives* (p-value < 0.001).

Subsequently, we conduct the same regressions but only including Round 1 Others in treatment *Primitives*. In this case, the treatment dummy is not statistically significant (p-value 0.130) and the coefficient is much smaller at 5.4 percentage points. If, instead, we include only subjects classified as Round 1 pBRN in *Primitives*, the treatment dummy is significant at the one percent level and the coefficient is larger at 14.9 percentage points.[33]

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.[34] 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. Separately looking at Round 1 pBRN subjects and Round 1 Others in *Primitives*, we find that the convergence rate is 83 percent (94 percent) by the last 10 (5) rounds for the Round 1 pBRN subjects and 68

[32]For details, see Table 7 of Online Appendix C.

[33]Focusing only on those subjects in *Primitives*, there is also evidence that subjects classified as Round 1 pBRN are significantly different from others (Round 1 Others). If we run a regression that only includes subjects in *Primitives*, where the right-hand side dummy takes value 1 if the subject is classified as Round 1 pBRN and zero otherwise, the coefficient on the treatment dummy equals 9.4 percentage points and is significant (p-value 0.041).

[34]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.

percent (93 percent) by the last 10 (5) rounds for the Round 1 Others.

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 .001). Within participants in *Primitives*, there is no difference between Round 1 pBRN subjects and Round 1 Others.

Discussion

The evidence suggests that there are large differences between subjects in *NoPrimitives* and *Primitives* in terms of how they make use of the feedback. Subjects in *NoPrimitives* are more responsive to immediate and cumulative feedback, are less likely to have converged after 100 rounds, and take more time to make decisions as they receive feedback. These differences are stable, and often amplified, when we contrast subjects in treatment *NoPrimitives* to only the Round 1 pBRN subjects in *Primitives*.[35] Our interpretation of these findings is that subjects in *Primitives*—particularly those classified as Round 1 pBRN—are less attentive to feedback.

It is useful at this point to discuss further what we mean by *attentiveness*. First, the experiment was designed such that the feedback was visually available to the subjects at any point at almost no cost. For example, the outcome of each round was prominently presented and the subjects were required to click on buttons on the same screen to proceed to the next round. Moreover, the outcome of each round was automatically recorded and displayed in a history table in all future rounds. While in principle subjects may actively try not to observe portions of their screens, these design features were put in place to minimize the costs associated with seeing and keeping track of the data. With attentiveness, we mean to capture a more meaningful notion in which subjects are not just looking at the data but are also engaging with it in a way that could challenge their beliefs. Note that given the stochastic nature of the task no single round of feedback can invalidate a subject's beliefs. Learning from feedback requires subjects to process the feedback in way that generates a compelling test of their model of the world. For example, looking at the empirical distribution of the state conditional on each

[35] Behavior of Round 1 Others in *Primitives* often lies in between the behavior of the other two groups: subjects in *NoPrimitives* and Round 1 pBRN subjects in *Primitives*. Overall, these observations are consistent with Round 1 Others in *Primitives* representing a collection of subjects who may adopt very different approaches to form beliefs (using both the feedback and the primitives). For example, this group includes subjects who submit Bayesian beliefs from round 1 (who may not be attentive to data) as well those subjects who do not initially have Bayesian or pBRN beliefs and follow a frequentist approach (who would closely track the data).

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 4: 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 .001 and not reported.

signal after 100 rounds provides a strong statistical argument that the pBRN model cannot be correct. While this information is readily available, subjects might choose not to engage with the data—potentially because their incorrect mental model endows no value to such an exercise—and hence fail to observe this pattern. This is precisely the type of inattentiveness we hope to capture in the experiment.

The results presented so far are consistent with the idea that long-run beliefs are farther from the Bayesian benchmark in *Primitives* because subjects who form incorrect mental models in this treatment (Round 1 pBRN subjects) fail to engage with the data. In the following sections, we provide further evidence for this idea by showing that (1) these subjects indeed have a noisier recollection of the feedback they experienced (suggesting that they do not process the data), and (2) beliefs of these subjects adjust substantially when feedback they have already experienced is summarized to them in a way that clearly challenges their model of the world.

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 Section 2, 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 4a, 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.[36] 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 the beliefs they submit.[37]

To study more carefully how well subjects recall feedback and how that connects to the beliefs they submit, in Table 4b 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 $j \in \{\text{Neg}, \text{Pos}\}$.[38] 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 4b). 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 4b).

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

Summary

The overall evidence in this section suggests that Round 1 pBRN subjects are not closely tracking and using feedback to inform their decisions relative to other subjects, and we summarize this finding below.

[36] Online Appendix D presents a more detailed analysis of each of the four reported values.

[37] 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.

[38] 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*).

Finding #3: *The evidence suggests that those subjects in Primitives who perfectly neglect the base rate in round 1 are less attentive to feedback.*

We revisit the estimation strategy introduced in Section 3.2 based on Grether (1980) to provide another perspective on how Round 1 pBRN subjects differ from others in the evolution of their responses. We estimate the parameters α (responsiveness to prior) and β (responsiveness to signal) separately for Round 1 pBRN subjects and Round 1 Others in treatment *Primitives*. The decomposition for all rounds is plotted in Figure 12 of Online Appendix C. After 200 rounds, the estimated value of α for Round 1 pBRN subjects is approximately 0.42, which is substantially lower than 0.71 estimated for Round 1 Others and 0.83 estimated for subjects in *NoPrimitives* (as well as the Bayesian benchmark of 1).

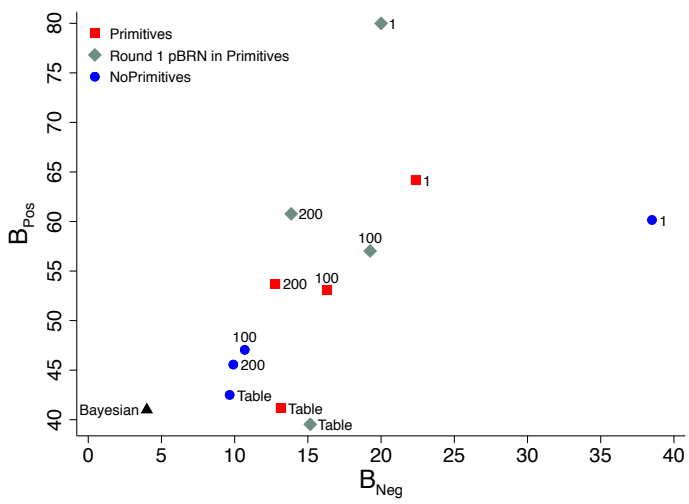
3.5 Summary tables

Preliminaries

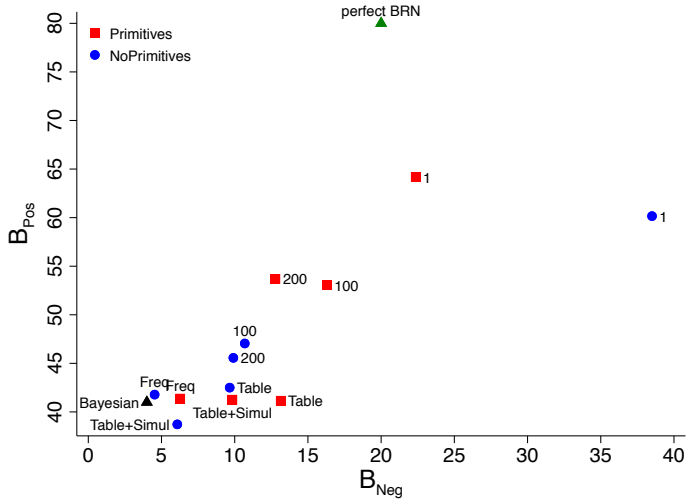
In this section we study the effect of showing subjects aggregate data (that they have already experienced) in a summarized table form. As explained in Section 2, we begin by presenting them with feedback from rounds 1-200 using a two-by-two table that reports the number of rounds that each of the four combinations of signal-state realizations were observed.[39]

We view the provision of the table as an intervention that significantly reduces the attention costs of the subjects. Recall that we view attention costs generally as the costs of using the data to confront one's model of the world. As the literature has pointed out, these costs include keeping track of the data and processing this data in an intelligent manner. Provision of the table may also decrease psychological costs of paying attention, such as the disutility of finding out one's initial beliefs were wrong, provided that it is costly to fool oneself in the presence of unequivocal evidence (e.g., Falk and Zimmermann (2018), Zimmermann (2018)). Hence, by providing subjects with a summary table, we are able to test directly the degree to which differential attention to feedback is driving the main results documented so far. Our main result in this section demonstrates that beliefs indeed cluster around the Bayesian benchmark (and move away from BRN) when feedback is presented in this form.

[39]Interventions where subjects are presented with aggregate information is common in the psychology literature. For example, Gigerenzer and 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.



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



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

Figure 7: 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.

Summary table: results

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 be observed in Figure 7a, 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*. The adjustment is even larger (approximately 22 percentage points) for Round 1 pBRN subjects. Moreover, as shown in Table 9 of Online Appendix C, once the feedback is summarized in this table form, differences in behavior between treatments *Primitives* and *NoPrimitives* disappear.[40] These patterns are also confirmed when we look at the aggregate degree of partial base rate neglect (following the estimation strategy introduced in 3.2 based on Grether (1980)). In *Primitives*, the estimate for α (which captures responsive-

[40]Beliefs in *Primitives* are not statistically farther away from the Bayesian benchmark relative beliefs in *NoPrimitives* (p-value .394) and the beliefs in *Primitives* are not statistically different from beliefs in *NoPrimitives* (p-value .523). If distance is measured relative to each subject's realized frequencies in their own data, we find the same qualitative result (Table 7 in Online Appendix C).

ness to prior) jumps from slightly above $.4$ to above $.8$ after subjects are presented with the table. Moreover, we can now reject the hypothesis that the estimates for α are different across treatments (p-value .262).

Discussion

The impact of the summary table, in terms of eliminating differences across treatments and the large adjustment in behavior that we observe from Round 1 pBRN subjects, suggests that incorrect mental models can be corrected provided that information that could initiate such a correction is presented to subjects in the right way. For other groups of subjects, the effect of the table is muted, providing further evidence that these subjects were already engaged with the data.

Contrasting behavior before and after the summary table allows us to estimate a measure of attentiveness. We estimate attentiveness by using a learning model as described in Online Appendix E, where, by considering subjects' beliefs in rounds 1-200 and in the part of the experiment where the table is provided, we recover an attention parameter that equals 0 (1) for completely inattentive (attentive) subjects. The estimate for Round 1 pBRN subjects equals 0.17, while it is at 0.98 and 1 for Round 1 Others and subjects in *NoPrimitives*, respectively.[41]

We summarize the findings from this section next:

Finding #4: *Summarizing feedback in table form has a significant impact on beliefs (especially on subjects in Primitives) and eliminates treatment effects.*

Additional data summaries

As explained in Section 2, 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 be-

[41] For example, an estimate of 0.17 for the attention parameter can be interpreted as subject internalizing only 17% of the observations they experience in updating their beliefs.

liefs, 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 in Online Appendix C. The distance between elicited beliefs and the Bayesian benchmark decreases from 25.3 (percentage points) in round 1, to 9.8 in phase 1 of the summary table, to 7.0 in phase 2, and to 2.5 in the final phase for the *Primitives* treatment. For *NoPrimitives*, this figure decreases from 27 (percentage points) in round 1, to 8.0 in phase 1 of the summary table, to 4.9 in phase 2, and to 2.1 in the final phase.[42] Moreover, as Figure 14 and Figure 15 in Online Appendix C show, essentially all subjects in both treatments report beliefs very close to the Bayesian benchmark by the end of the final phase.[43]

Finally, we summarize our findings after the final phase of data summaries:

Finding #5: *After subjects are directly presented with realized frequencies, almost all subjects in both treatments report beliefs very close to the Bayesian benchmark.*

3.6 Transfer learning

Preliminaries

In treatment *Primitives*, the average belief conditional on a positive signal (B_{Pos}) moves from 64 percent in round 1 to the Bayesian value of 41 percent after 200 rounds of feedback and the provision of summary tables. Clearly, the feedback is instrumental in shaping subjects' beliefs, but it is natural to ask exactly what subjects learn from this exercise and whether learning can be transferred to a new setting. In particular, have subjects learned that their mental model was incorrect because it neglected the information on the prior?[44]

We tackle this question in the last part of the experiment, where subjects face a new updating task in which the primitives are changed to $p' = .95$ and $q' = .85$. Subjects are asked to report beliefs just once, without any feedback. We call this final round with new values p' and

[42] See Table of 9 of Online Appendix C for details.

[43] By the final phase, where a table with frequencies is provided, 96 and 95 percent of subjects submit beliefs within ± 5 percentage points of the realized frequencies in treatments *Primitives* and *NoPrimitives*, respectively.

[44] A few papers have studied transfer of learning across environments and find limited evidence for it (e.g. Kagel (1995), Cooper and Kagel (2009), Cooper and Van Huyck (2018)).

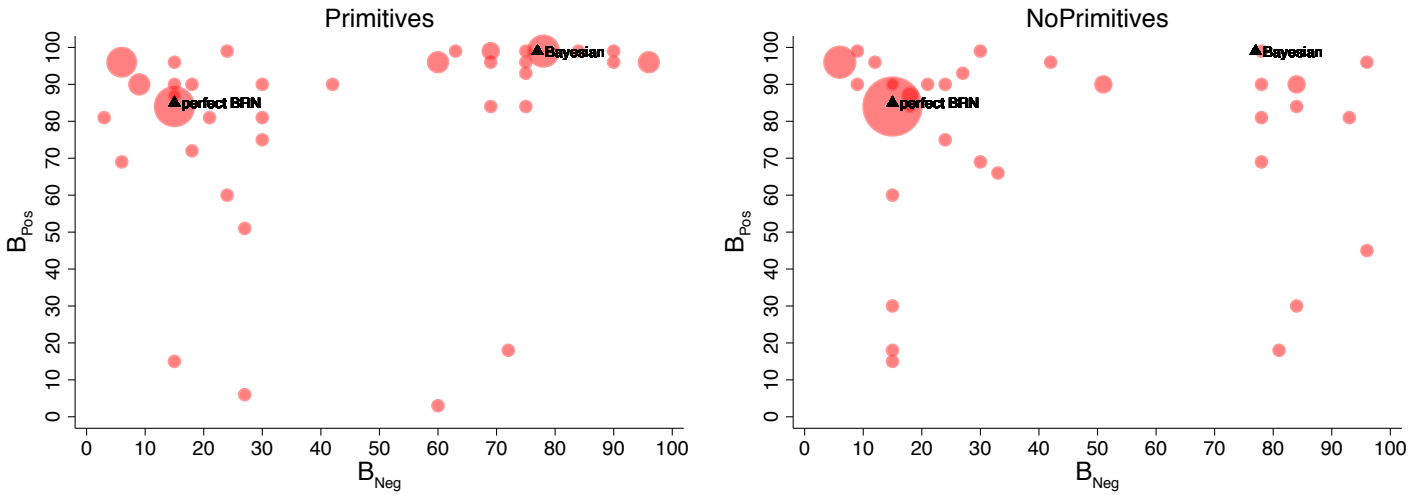


Figure 8: Transfer learning: density plots in final round with new primitives

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 where the prior and the reliability of the signal is changed.

q' "round$_{(p',q')}$" and assess how beliefs differ from the Bayesian benchmark relative to answers in round 1 and an appropriate control, as described below.

Results

On average, beliefs in *Primitives* for round$_{(p',q')}$ equal $(41, 85)$, where the benchmarks are $(B^{*'}_{Neg}, B^{*'}_{Pos}) = (77, 99)$ for Bayesian updating and $(B^{pBRN'}_{Neg}, B^{pBRN'}_{Pos}) = (15, 85)$ for perfect base-rate neglect, pBRN. But, as the distribution of responses plotted in Figure 8 reveals, there are essentially two clusters of responses; one cluster is close to pBRN and the other to the Bayesian benchmark. A comparison with the left plot of Figure 4 suggests that subjects are providing beliefs closer to the Bayesian posterior in round$_{(p',q')}$ relative to round 1. Specifically, 17.2 percent of subjects provide pBRN beliefs in this part, a figure that is below the 56.2 percent observed in round 1. Meanwhile, 12.5 percent of subjects provide beliefs exactly at the Bayesian benchmark compared to 4.7 percent in round 1.

However, it is difficult to interpret the comparison between round 1 and round$_{(p',q')}$ in *Primitives* without an appropriate control. We use the *NoPrimitives* treatment as an effective control. In particular, subjects in this treatment were also told the primitives $p' = .95$ and $q' = .85$ for round$_{(p',q')}$, and so subjects in both of our treatments faced this final round in an

identical manner. Because only subjects in *Primitives* could be induced to have the base-rate neglect mental model in rounds 1-200, if there is any correction to the mental model that is transferable across settings, then one should expect these subjects to incorporate the prior more in round$_{(p',q')}$ relative to subjects in the *NoPrimitives* treatment, who also had an opportunity to learn from the data but who could not plausibly hold the incorrect mental model in rounds 1-200. In other words, since subjects in *NoPrimitives* did not know the primitives in rounds 1-200, it is not possible for them to have learned a general lesson on how the base-rate should not be neglected in round$_{(p',q')}$, where the prior and accuracy are now given to them.

Average beliefs in round$_{(p',q')}$ for *NoPrimitives* equal $(30, 81)$. That is, while the average for B_{Pos} is similar across treatments, there is a significant 11 percentage point difference in terms of B_{Neg} (p-value .028), which is the dimension on which the Bayesian and pBRN benchmarks differ the most, with the belief being closer to the Bayesian benchmark in *Primitives*. However, the difference is more clearly visible in the distribution of beliefs across treatments as presented in Figure 8. In *NoPrimitives* there are almost no subjects around the Bayesian benchmark (1.6 percent), but a relatively larger group is concentrated around the pBRN point (37.5 percent). In fact, if we allow for $\pm$ 5 percentage points in each belief, then 47 percent of subjects in *NoPrimitives* and 25 percent of subjects in *Primitives* are classified as pBRN. Meanwhile, a similar exercise does not change the proportion of subjects submitting Bayesian beliefs in *NoPrimitives* (still 1.6), but it increases to 15.6 in *Primitives*.

This treatment effect (which switches direction relative to earlier parts!) suggests that at least some subjects in *Primitives* can extrapolate from what they learned with the baseline primitives to new primitives.[45] However, we should also note that such learning is partial as average beliefs continue to be far from the Bayesian benchmark.

Finding #6: *When subjects are exposed to a new environment with different but known primitives, the treatment effect reverses: Subjects in NoPrimitives now neglect the prior to a significantly larger extent than subjects in Primitives. This suggests that some subjects in Primitives can learn to take the prior into account when facing a new environment.*

[45] Using the estimation strategy introduced in Section 3.2 based on Grether (1980), the estimate for α (responsiveness to prior) increases from .21 in round 1 to .31 in round$_{(p',q')}$ for *Primitives*. Meanwhile, the corresponding estimate in round$_{(p',q')}$ is .13. for *NoPrimitives*.

4 Conclusion

Behavioral biases—particularly in initial responses—are well documented in the literature. However, much less is known on whether such biases are persistent in the long run in information-rich environments where agents have opportunities to correct these biases. In this paper, we focus on a simple updating task and study whether base-rate neglect (BRN, one of the most robustly documented biases in the literature) is persistent in the long run in the presence of feedback and assess the role that incorrect mental models (such as BRN) play in hindering learning from feedback. To provide insights on this, we compare beliefs in a baseline treatment, in which a majority of subjects display base-rate neglect in initial beliefs, to a control treatment that does not allow for BRN as a mental model but in which learning from feedback is similarly possible. We achieve this by providing the primitives of the updating task to the subjects in the baseline treatment but not in the control treatment.

While we document evidence of learning from feedback, adjustment of beliefs in response to feedback is slow and partial in the baseline treatment. After 200 rounds of feedback, beliefs are farther from the Bayesian benchmark in the baseline relative to the control treatment. We also document that the treatment difference is driven by those subjects who start with an incorrect mental model (BRN) in the baseline. In addition, the limited degree of learning displayed by these subjects is linked to partial attentiveness to feedback. These subjects are less responsive to the feedback and have a noisier recollection of the outcomes they have experienced. Yet we also show that these subjects are able to correct their beliefs in response to unequivocal evidence that goes against their mental model: When we lower attention costs substantially, by summarizing feedback in a table form, BRN largely disappears. Finally, we also find some (but limited) evidence that learning from feedback can generate insights (for example, that the base rate should be considered in the belief formation process) that can be transferred to a new setting.

In conclusion, our results demonstrate how suboptimal behavior can be persistent in the long run even in information rich environments. Initial misconceptions, which drive suboptimal behavior, can also prevent learning from feedback by impacting an agent's attentiveness to this information. This insight also connects closely with the literature on learning with misspecified models and learning with endogenous attention, as we discuss in detail in the introduction. An important implication of our results is that for interventions designed to counter systematic biases to succeed, they need to move beyond providing information that is indicative of optimal behavior and target agents engagement with this information. We also find that

withholding information that agents consider as payoff-relevant can increase attentiveness to feedback and foster learning.

We see several directions in which this research agenda can be advanced. First, our paper focuses on base-rate neglect as a proof of concept, but the experimental design we propose can be incorporated to other settings to study the persistence of other well-documented biases such as overconfidence or correlation neglect in response to different forms of feedback. Adopting this approach more broadly can help better identify what types of biases are persistent even in information rich environments. Second, there are other channels through which initial misconceptions can prevent learning from feedback. We focus on a simple decision problem where feedback was exogenous to an agent's decision. Moving beyond this paradigm—looking at games and decision problems with endogenous feedback—would uncover other forces that contribute to the persistence of suboptimal behavior. Finally, while the controlled environment the laboratory provides is a natural starting point to study the interaction between biases and learning, we believe that it is important to assess the extent to which biases persist in prominent field applications. For example, future work can study base-rate neglect in doctors interpreting medical tests using types of feedback that are natural in that setting.

References

AGRANOV, M., U. DASGUPTA, AND A. SCHOTTER (2018): "Trust Me: Communication and Competition in Psychological Games," *Working Paper*.

ARAUJO, F., S. WANG, AND A. WILSON (2019): "The times they are a-Changing: Dynamic Adverse Selection in the Laboratory," *Working Paper*.

ARKES, H. R. AND C. BLUMER (1985): "The psychology of sunk cost," *Organizational behavior and human decision processes*, 35, 124–140.

BAR-HILLEL, M. (1980): "The base-rate fallacy in probability judgments," *Acta Psychologica*, 44, 211–233.

BARBEY, A. K. AND S. A. SLOMAN (2007): "Base-rate respect: From ecological rationality to dual processes," *Behavioral and Brain Sciences*, 30, 241–254.

BARRON, K., S. HUCK, AND P. JEHIEL (2019): "Everyday econometricians: Selection neglect and overoptimism when learning from others," *Working Paper*.

BAYONA, A., J. BRANDTS, AND X. VIVES (2020): "Information Frictions and Market Power: A Laboratory Study," *Games and Economic Behavior*.

BÉNABOU, R. AND J. TIROLE (2003): "Intrinsic and extrinsic motivation," *The review of economic studies*, 70, 489–520.

——— (2016): "Mindful economics: The production, consumption, and value of beliefs," *Journal of Economic Perspectives*, 30, 141–64.

BENJAMIN, D., A. BODOH-CREED, AND M. RABIN (2019): "Base-Rate Neglect: Foundations and Implications," *Working Paper*.

BENJAMIN, D. J. (2019): "Errors in probabilistic reasoning and judgment biases," 2, 69–186.

BOHREN, J. A. AND D. N. HAUSER (2017): "Bounded rationality and learning: A framework and a robustness result," *Working Paper*.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2013): "Salience and consumer choice," *Journal of Political Economy*, 121, 803–843.

BRUNNERMEIER, M. K. AND J. A. PARKER (2005): "Optimal expectations," *American Economic Review*, 95, 1092–1118.

CAPLIN, A. AND M. DEAN (2015): "Revealed preference, rational inattention, and costly information acquisition," *American Economic Review*, 105, 2183–2203.

CASON, T. N. AND C. R. PLOTT (2014): "Misconceptions and game form recognition: Challenges to theories of revealed preference and framing," *Journal of Political Economy*, 122, 1235–1270.

CHARNESS, G., R. OPREA, AND S. YUKSEL (2019): "How do people choose between biased information sources? Evidence from a laboratory experiment," *Working Paper*.

CHRISTENSEN-SZALANSKI, J. J. AND L. R. BEACH (1982): "Experience and the base-rate fallacy," *Organizational Behavior and Human Performance*, 29, 270–278.

CIPRIANI, M. AND A. GUARINO (2009): "Herd behavior in financial markets: an experiment with financial market professionals," *Journal of the European Economic Association*, 7, 206–233.

COOPER, D. J. AND J. H. KAGEL (2009): "The role of context and team play in cross-game learning," *Journal of the European Economic Association*, 7, 1101–1139.

COOPER, D. J. AND J. VAN HUYCK (2018): "Coordination and transfer," *Experimental Economics*, 21, 487–512.

COSMIDES, L. AND J. TOOBY (1996): "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty," *cognition*, 58, 1–73.

COX, J. C. AND R. L. OAXACA (2000): "Good news and bad news: Search from unknown wage offer distributions," *Experimental Economics*, 2, 197–225.

DAL BÓ, E., P. DAL BÓ, AND E. EYSTER (2018): "The demand for bad policy when voters underappreciate equilibrium effects," *The Review of Economic Studies*, 85, 964–998.

DANZ, D., L. VESTERLUND, AND A. J. WILSON (2020): "Belief elicitation: Limiting truth telling with information on incentives," *National Bureau of Economic Research*.

DEKEL, E., D. FUDENBERG, AND D. LEVINE (2004): "Learning to play Bayesian games," *Games and Economic Behavior*, 46, 282–303.

ENKE, B. (2019): "What you see is all there is," *Working Paper*.

ENKE, B. AND F. ZIMMERMANN (2019): "Correlation neglect in belief formation," *The Review of Economic Studies*, 86, 313–332.

ESPONDA, I. AND D. POUZO (2016): "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 84, 1093–1130.

ESPONDA, I. AND E. VESPA (2014): "Hypothetical Thinking and Information Extraction in the Laboratory," *American Economic Journal: Microeconomics*, 6, 180–202.

——— (2018): "Endogenous sample selection: A laboratory study," *Quantitative Economics*, 9, 183–216.

——— (2019): "Contingent Thinking and the Sure-Thing Principle: Revisiting Classic Anomalies in the Laboratory," *Working paper*.

EYSTER, E. AND G. WEIZSÄCKER (2010): "Correlation neglect in financial decision-making," *Working Paper*.

FALK, A. AND F. ZIMMERMANN (2018): "Information processing and commitment," *The Economic Journal*, 128, 1983–2002.

FANTINO, E. AND A. NAVARRO (2012): "Description–experience gaps: Assessments in other choice paradigms," *Journal of Behavioral Decision Making*, 25, 303–314.

FISCHBACHER, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

FUDENBERG, D., G. ROMANYUK, AND P. STRACK (2017): "Active learning with a misspecified prior," *Theoretical Economics*, 12, 1155–1189.

FUDENBERG, D. AND E. VESPA (2019): "Learning Theory and Heterogeneous Play in a Signaling-Game Experiment," *American Economic Journal: Microeconomics*, 11, 186–215.

GABAIX, X. (2014): "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 129, 1661–1710.

GAGNON-BARTSCH, T., M. RABIN, AND J. SCHWARTZSTEIN (2018): "Channeled attention and stable errors," *Working paper*.

GENNAIOLI, N. AND A. SHLEIFER (2010): "What comes to mind," *The Quarterly journal of economics*, 125, 1399–1433.

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

GIGERENZER, G. AND U. HOFFRAGE (1995): "How to improve Bayesian reasoning without instruction: frequency formats." *Psychological review*, 102, 684.

GOODIE, A. S. AND E. FANTINO (1996): "Learning to commit or avoid the base-rate error," *Nature*, 380, 247.

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

GRAEBER, T. (2019): "Inattentive inference," *Working Paper*.

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

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

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

GRIFFIN, D. AND A. TVERSKY (1992): "The weighing of evidence and the determinants of confidence," *Cognitive psychology*, 24, 411–435.

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

HANNA, R., S. MULLAINATHAN, AND J. SCHWARTZSTEIN (2014): "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 129, 1311–1353.

HEIDHUES, P., B. KŐSZEGI, AND P. STRACK (2018): "Unrealistic expectations and misguided learning," *Econometrica*, 86, 1159–1214.

HUCK, S., P. JEHIEL, AND T. RUTTER (2011): "Feedback spillover and analogy-based expectations: A multi-game experiment," *Games and Economic Behavior*, 71, 351–365.

HUFFMAN, D., C. RAYMOND, AND J. SHVETS (2018): "Persistent Overconfidence and Biased Memory: Evidence from Managers," *Working paper*.

JOHNSON-LAIRD, P. N. (1980): "Mental models in cognitive science," *Cognitive science*, 4, 71–115.

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

KAHNEMAN, D. AND A. TVERSKY (1972): "On prediction and judgement," *ORI Research Monograph*, 12.

——— (1973): "On the psychology of prediction." *Psychological review*, 80, 237.

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

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

LINDEMAN, S. T., W. P. VAN DEN BRINK, AND J. HOOGSTRATEN (1988): "Effect of feedback on base-rate utilization," *Perceptual and Motor Skills*, 67, 343–350.

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

MANIS, M., I. DOVALINA, N. E. AVIS, AND S. CARDOZE (1980): "Base rates can affect individual predictions." *Journal of Personality and Social Psychology*, 38, 231.

MARTIN, D. AND E. MUÑOZ-RODRIGUEZ (2019): "Misperceiving Mechanisms: Imperfect Perception and the Failure to Recognize Dominant Strategies," *Working paper*.

MARTÍNEZ-MARQUINA, A., M. NIEDERLE, AND E. VESPA (2019): "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*, 109, 3437–74.

MEDIN, D. L. AND S. M. EDELSON (1988): "Problem structure and the use of base-rate information from experience." *Journal of Experimental Psychology: General*, 117, 68.

MOBIUS, M. M., M. NIEDERLE, P. NIEHAUS, AND T. S. ROSENBLAT (2011): "Managing self-confidence: Theory and experimental evidence," *Working paper*.

MOORE, D. A. AND P. J. HEALY (2008): "The trouble with overconfidence." *Psychological review*, 115, 502.

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

NGANGOUE, K. AND G. WEIZSÄCKER (2018): "Learning from Unrealized versus Realized Prices," *Working paper*.

NISBETT, R. E., E. BORGIDA, R. CRANDALL, AND H. REED (1976): "Popular induction: Information is not necessarily informative," .

RABIN, M. (2000): "Inference by believers in the law of small numbers," *Quarterly journal of Economics*, 117, 775–816.

SCHOTTER, A. AND Y. M. BRAUNSTEIN (1981): "Economic search: an experimental study," *Economic inquiry*, 19, 1–25.

SCHWARTZSTEIN, J. (2014): "Selective attention and learning," *Journal of the European Economic Association*, 12, 1423–1452.

SIMS, C. A. (2003): "Implications of rational inattention," *Journal of monetary Economics*, 50, 665–690.

THALER, R. (1980): "Toward a positive theory of consumer choice," *Journal of Economic Behavior & Organization*, 1, 39–60.

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

ZIMMERMANN, F. (2018): "The dynamics of motivated beliefs," *American Economic Review*.

ZUKIER, H. AND A. PEPITONE (1984): "Social roles and strategies in prediction: Some determinants of the use of base-rate information." *Journal of Personality and Social Psychology*, 47, 349.

ONLINE APPENDIX FOR

MENTAL MODELS AND LEARNING:

THE CASE OF BASE-RATE NEGLECT

Ignacio Esponda Emanuel Vespa Sevgi Yuksel

CONTENTS:

A. Literature Review on BRN with feedback

B. Experimental Instructions

C. Additional Tables and Figures

D. Additional Analysis of Recollection of Feedback

E. Estimates from a Learning Model

A Literature Review on BRN with feedback

This section 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.

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.[46] 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 and 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.[47] 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 and Tversky, 1972) or experiments where either the prior or the signal reliability is open to interpretation (as in Kahneman and Tversky, 1973).

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

[46] 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).

[47] 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.

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 further evidence on different aspects of this are also presented in Cosmides and 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.[48] 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.[49]

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

[48] 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).

[49] 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.

the primitives determining statistical relationship between the cue and the optimal action.[50] 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.[51] 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 and 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.

B 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

[50] 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).

[51] 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 and 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.

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):

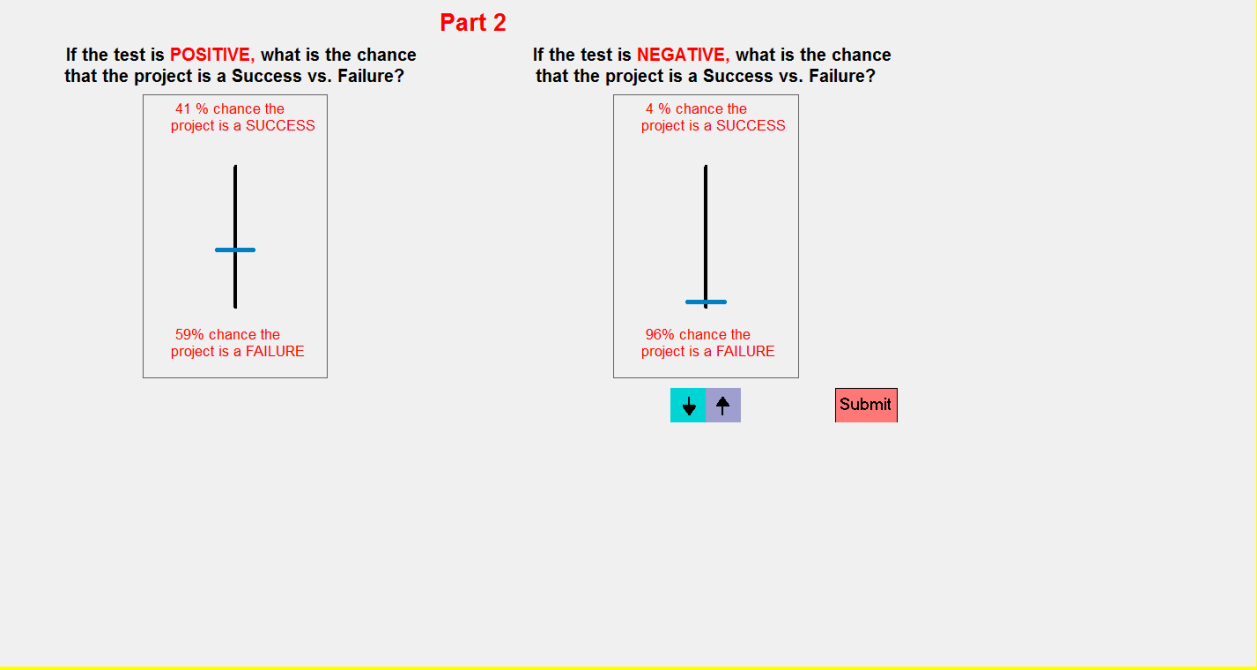


Figure 9: Interface screenshots for round 1 (presented as part 2 to subjects)

C Additional Tables and Figures

Sample	(1) R1 pBRN v. NoP		(2) $B_{Pos} \geq 70, B_{Pos} \leq 30$		(3) R1 pBRN v. R1 Others		(4) R1 Others v. NoP	
	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}	γ_{Pos}	γ_{Neg}
R1	19.8	-18.5	2.0	0.4	36.2	-5.4	-16.3	-13.1
	(.000)	(.000)	(.186)	(.790)	(.000)	(.192)	(.000)	(.002)
R100	10.0	8.6	12.5	11.6	9.1	6.8	0.9	1.8
	(.047)	(.010)	(.079)	(.015)	(.157)	(.115)	(.858)	(.486)
R200	15.2	3.9	15.5	6.4	16.2	2.5	-0.9	1.5
	(.000)	(.068)	(.012)	(.008)	(.003)	(.279)	(.808)	(.539)
#Obs	100		60		64		92	

Table 5: Estimation output for subsets of subjects

Notes: The table presents different estimates of γ_{Pos} and γ_{Neg}, where $B_{Pos} = \delta_{Pos} + \gamma_{Pos}\text{D} + \upsilon_{Pos}$ and $B_{Neg} = \delta_{Neg} + \gamma_{Neg}\text{D} + \upsilon_{Neg}$. Equations are estimated jointly using the seemingly unrelated regressions procedure. Each row constrains the sample to the decision referred to in the first column, where R refers to a round. In (1) the dummy D takes value 1 if the subject was classified as Round 1 pBRN in *Primitives* and 0 if the subject participated in *NoPrimitives*. In (2) D takes value 1 if the subject is in Primitives and as 0 if in NoP, but the sample is restricted to subjects who in round 1 (R1) submitted beliefs such that: $B_{Pos} \geq 70$ and $B_{Neg} \leq 30$. The was classified as Round 1 pBRN in *Primitives* and 0 if the subject participated in *NoPrimitives*. In (3) the dummy D 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 Round 1 Others in *Primitives*). In (4) dummy D takes value 1 if the subject is classified as 'Round 1 Others in *Primitives*' and 0 if the subject participated in the *NoPrimitives*. Between parentheses we report standard errors. The last row indicates the number of observations in each regression.

Sample	(1) Dep. var: Δ (distance to benchmark)				(2) Dep. vars: B_{Neg}, B_{Pos}				
	a		b		γ_{Neg}		γ_{Pos}		H_0:
	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	$\gamma_{Neg} = \gamma_{Pos} = 0$
Round 1	27.0	.000	-1.7	.203	-16.1	.000	4.0	.258	.000
Round 20	20.9	.000	0.4	.847	-0.4	.889	2.6	.560	.829
Round 100	13.0	.000	5.6	.003	5.6	.035	6.0	.159	.056
Round 200	10.5	.000	4.8	.004	2.9	.112	8.1	.021	.049
Table - 200	8.0	.000	1.8	.394	3.5	.346	-1.3	.549	.523
Table - 1000	4.9	.000	2.1	.140	3.8	.053	2.5	.204	.089
Table - 1000 - freq	2.1	.002	0.4	.691	1.8	.242	-0.4	.692	.453

Table 6: Estimation output

Notes: Each row presents the results for two sets of regressions. Columns under (1) report the estimates (coeff. column) and p-values (for the null that the corresponding estimate is zero) of: $\Delta = a + b\text{P} + \epsilon$, where ϵ is a noise parameter and P a treatment dummy (1 if the observation belongs to *Primitives*). Columns under (2) report the γ estimates and corresponding p-values for: $B_{Neg} = \delta_{Neg} + \gamma_{Neg}\text{P} + \upsilon_{Neg}$ and $B_{Pos} = \delta_{Pos} + \gamma_{Pos}\text{P} + \upsilon_{Pos}$, which are estimated jointly using the seemingly unrelated regressions procedure. The last column reports a Wald test in which the null hypothesis is that $\gamma_{Neg} = \gamma_{Pos} = 0$. Each row constrains the sample to the beliefs referred to in the first column, separating those parts where the feedback is presented in table form (Table - 200 refers to when subjects are presented with the summary table of the first 200 rounds, Table - 1000 refers to when subjects are presented with the table including 800 additional simulated rounds and Table- 1000 - freq refers to when subjects are presented with a table where the relevant frequencies are provided). Each regression involves 128 observations (64 from each treatment).

	Distance relative to Bayes' rule				Distance relative to frequencies			
	a		b		a		b	
	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value
Round 1	27.0	.000	-1.7	.203		-		
Round 20	20.9	.000	0.4	.847	26.8	.000	11.2	.004
Round 100	13.0	.000	5.6	.003	18.9	.000	10.8	.000
Round 200	10.5	.000	4.8	.004	16.0	.000	9.4	.001
Table - 200	8.0	.000	1.8	.394	8.9	.000	3.7	.146
Table - 1000	4.9	.000	2.1	.140	7.6	.000	2.8	.183
Table - 1000 - freq	2.1	.002	.4	.691	2.1	.019	0.3	.808

Table 7: Distance regressions: outputs using average absolute distance relative to Bayes rule and relative to realized frequencies

Notes: Each row presents the estimates (coeff. column) and p-values of: $\Delta = a + b\text{P} + \epsilon$, where ϵ is a noise parameter and P a treatment dummy (1 if the observation belongs to *Primitives*) for two different distance measures. The distance variable Δ is measured using the Bayesian predictions (realized frequencies) as a benchmark in the columns entitled 'Distance relative to Bayes' rule' ('Distance relative to realized frequencies'). Each row constrains the sample to the decision referred to in the first column and each regression involves 128 observations (64 from each treatment). The distance relative to Bayes' rule colums are reported in part (1) of Table 1 in the main body of the paper.

	Average absolute distance				Euclidean distance			
	a		b		a		b	
	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value
Round 1	27.0	.000	-1.7	.203	27.0	.000	-1.7	.203
Round 20	20.9	.000	0.4	.847	26.8	.000	11.2	.004
Round 100	13.0	.000	5.6	.003	18.9	.000	10.8	.000
Round 200	10.5	.000	4.8	.004	16.0	.000	9.4	.001
Table - 200	8.0	.000	1.8	.394	8.9	.000	3.7	.146
Table - 1000	4.9	.000	2.1	.140	7.6	.000	2.8	.183
Table - 1000 - freq	2.1	.002	.4	.691	2.1	.019	0.3	.808

Table 8: Distance regressions: outputs using absolute average distance to the Bayesian benchmark and euclidean distance to the Bayesian benchmark.

Notes: Each row presents the estimates (coeff. column) and p-values of: $\Delta = a + b\text{P} + \epsilon$, where ϵ is a noise parameter and P a treatment dummy (1 if the observation belongs to *Primitives*) for two different distance measures. The distance variable Δ is measured using the average absolute distance (Euclidean distance) relative to Bayesian predictions as a benchmark in the columns entitled 'Average absolute distance' ('Euclidean Distance'). Each row constrains the sample to the decision referred to in the first column and each regression involves 128 observations (64 from each treatment). The average absolute distance columns are reported in part (1) of Table 1 in the main body of the paper.

Sample	(1) Dep. var: Δ				(2) Dep. vars: B_{Neg}, B_{Pos}				
	a		b		γ_{Neg}		γ_{Pos}		H_0:
	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	Coeff.	p-value	$\gamma_{Neg} = \gamma_{Pos} = 0$
Round 1	27.4	.000	-1.7	.196	-15.9	.000	3.6	.297	.000
Round 20	19.2	.000	0.1	.952	-0.7	.806	2.2	.615	.850
Round 100	11.6	.000	5.3	.004	5.2	.041	5.7	.175	.064
Round 200	8.8	.000	4.6	.005	2.6	.123	7.9	.025	.055
Table - 200	6.1	.035	1.7	.421	3.3	.367	-1.2	.586	.561
Table - 1000	6.8	.001	2.0	.160	3.7	.057	2.7	.180	.088
Table - 1000 - freq	4.4	.001	0.4	.675	1.7	.232	-0.4	.702	.447

Table 9: Estimation output

Notes: Each row presents the results for two sets of regressions. Columns under (1) report the estimates (coeff. column) and p-values (for the null that the corresponding estimate is zero) of: $\Delta = a + b\text{P} + cX + \epsilon$, where X captures information collected in the survey ϵ is a noise parameter and P a treatment dummy (1 if the observation belongs to *Primitives*). Specifically, X includes three variables: 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. Columns under (2) report the γ estimates and corresponding p-values for: $B_{Neg} = \delta_{Neg} + \gamma_{Neg}\text{P} + \tau_{Neg}X + \upsilon_{Neg}$ and $B_{Pos} = \delta_{Pos} + \gamma_{Pos}\text{P} + \tau_{P}X + \upsilon_{Pos}$, which are estimated jointly using the seemingly unrelated regressions procedure. The last column reports a Wald test in which the null hypothesis is that $\gamma_{Neg} = \gamma_{Pos} = 0$. Each row constrains the sample to the decision referred to in the first column and each regression involves 128 observations (64 from each treatment).

	(1) Dep. Var.: $B^t_{Pos} - B^{t-1}_{Pos}$				(2) Dep. Var.: $B^t_{Neg} - B^{t-1}_{Neg}$			
	Primitives			*NoPrimitives*	*Primitives*			*NoPrimitives*
	All	Round 1 pBRN	Round 1 Others	All	All	Round 1 pBRN	Round 1 Others	All
(Pos, Succ)$_{t-1}$	0.7	0.5	1.0	2.4	0.1	0.2	-0.1	0.0
	(.028)	(.201)	(.076)	(.001)	(.770)	(.594)	(.377)	(.894)
(Pos, Fail)$_{t-1}$	-1.1	-0.9	-1.4	-2.8	0.1	0.1	0.1	0.4
	(.008)	(.147)	(.006)	(.000)	(.832)	(.922)	(.344)	(.072)
(Neg, Succ)$_{t-1}$	-0.2	-0.6	0.2	-0.9	0.9	0.6	1.3	2.0
	(.611)	(.476)	(.725)	(.069)	(.007)	(.149)	(.020)	(.004)
(Neg, Fail)$_{t-1}$	0.0	-.2	0.2	0.1	-0.2	-0.1	-0.3	-0.6
	(.951)	(.275)	(.023)	(.224)	(.045)	(.510)	(.004)	(.000)

Table 10: Changes in reported beliefs after feedback

Notes: Each column in group (1) and (2) provides estimates for regressions in which the dependent variable is the change in beliefs conditional on each signal (B_{Pos} in (1) and B_{Neg}in (2)) from round $t-1$ to t. The right-hand side consists of a set of dummies that cover all possible outcomes in the previous round. For example, (Pos, Succ)$_{t-1}$ is a dummy that takes value 1 if the feedback in the previous period was that the signal was positive and the state (the type of the project) was a success. The first three columns report results for subjects in *Primitives*. First including all subjects ('All'), then including only subjects who were classified as Round 1 pBRN and lastly including subjects who were not classified in round 1 as pBRN ('Round 1 Others'). The fourth column reports results for all subjects in *NoPrimitives*. Each regression includes data from all beliefs submitted in in rounds 2 to 100 (Part 3). Between parentheses we report p-values for the null hypothesis in which the coefficient equals zero. Standard errors used in the computation of the p-values are estimated by clustering at the subject level.

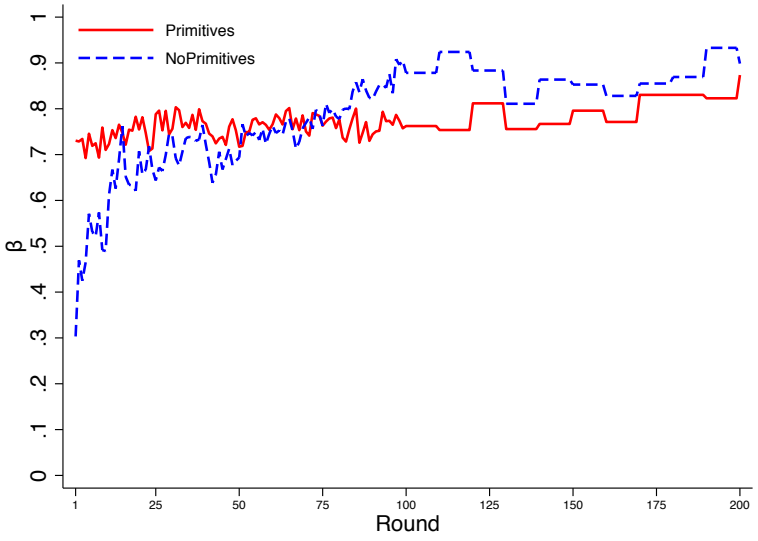


Figure 10: Estimate of β per round by treatment

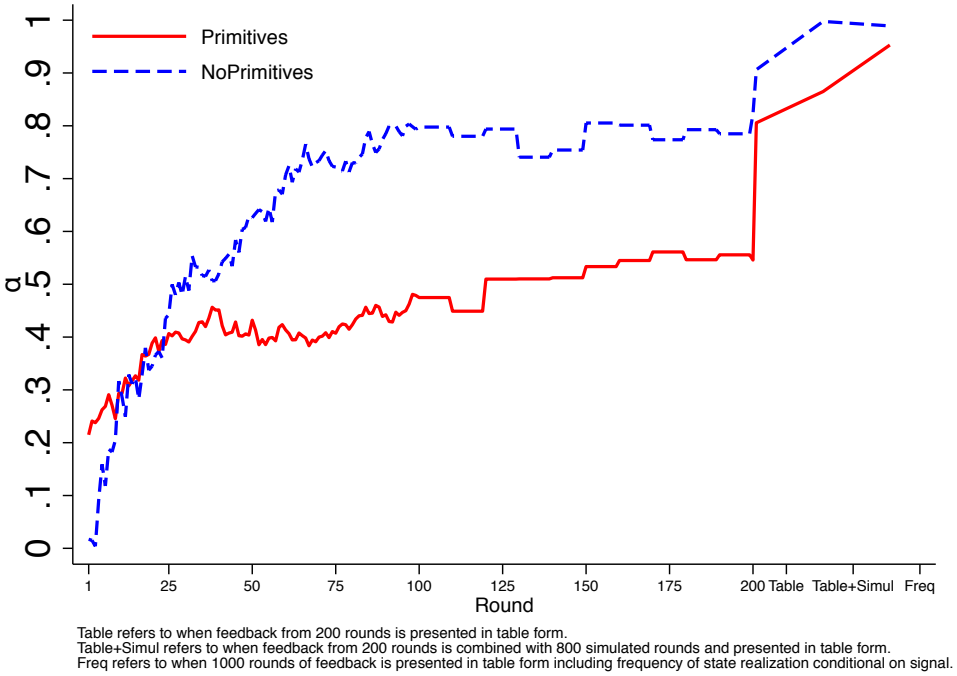


Figure 11: Estimate of α by treatment (parts 3 to 8)

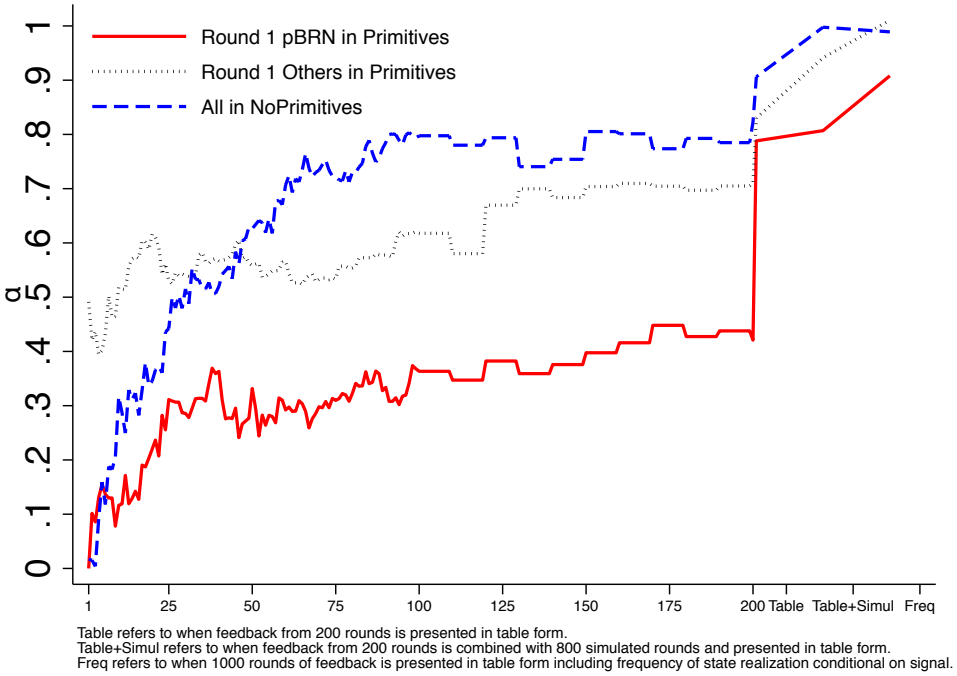


Figure 12: Estimate of α by treatment, where *Primitives* is decomposed by Round 1 pBRN and Round 1 Others

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 11: Estimation output for subsets of subjects

Notes: The table presents different estimates of γ_{Pos} and γ_{Neg}, where $B_{Pos} = \delta_{Pos} + \gamma_{Pos}\text{P} + \upsilon_{Pos}$ and $B_{Neg} = \delta_{Neg} + \gamma_{Neg}\text{P} + \upsilon_{Neg}$. 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.

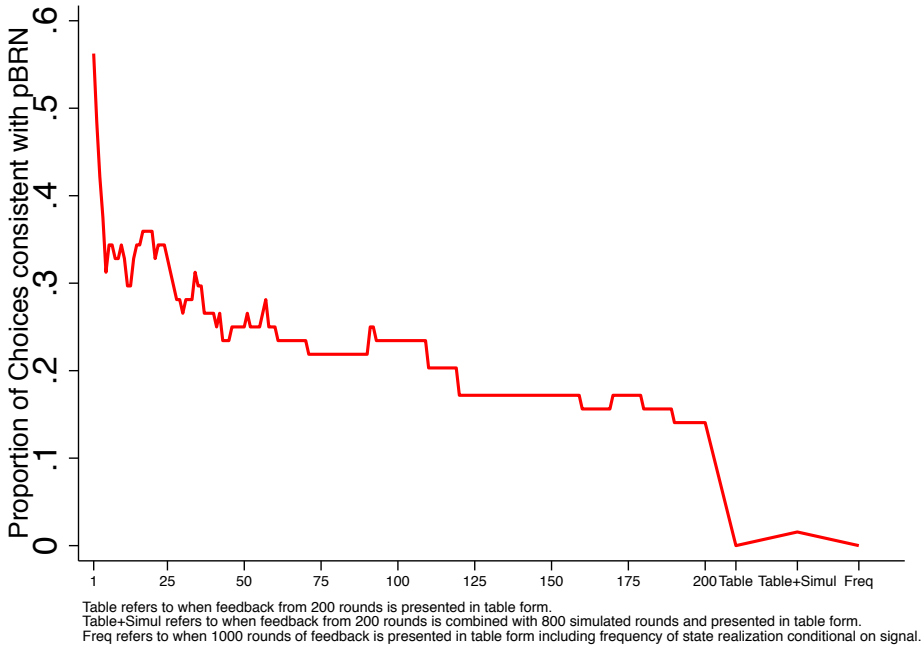


Figure 13: Proportion of choices consistent with pBRN in *Primitives* as the session evolves

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	-18.7	2.3	0.1	32.8	-6.5	-15.7	-11.9
	(.000)	(.000)	(.116)	(.941)	(.000)	(.104)	(.000)	(.003)
Round 100	9.9	7.6	13.6	10.9	10.2	5.2	0.8	2.1
	(.051)	(.019)	(.055)	(.020)	(.105)	(.219)	(.877)	(.411)
Round 200	15.3	3.9	15.2	6.5	14.8	2.0	-1.1	1.5
	(.000)	(.061)	(.002)	(.005)	(.006)	(.337)	(.779)	(.515)
Table -1000- freq	-0.4	3.3	0.9	2.4	0.4	3.2	-0.8	-0.3
	(.930)	(.091)	(.542)	(.464)	(.728)	(.268)	(.572)	(.461)
#Obs	100		60		64		92	

Table 12: Estimation output for subsets of subjects including survey controls

Notes: The table presents different estimates of γ_{Pos} and γ_{Neg}, where $B_{Pos} = \delta_{Pos} + \gamma_{Pos}\text{P} + \tau_{Pos}X + \upsilon_{Pos}$ and $B_{Neg} = \delta_{Neg} + \gamma_{Neg}\text{P} + \tau_{Neg}X + \upsilon_{Neg}$. X includes three variables: 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. 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.

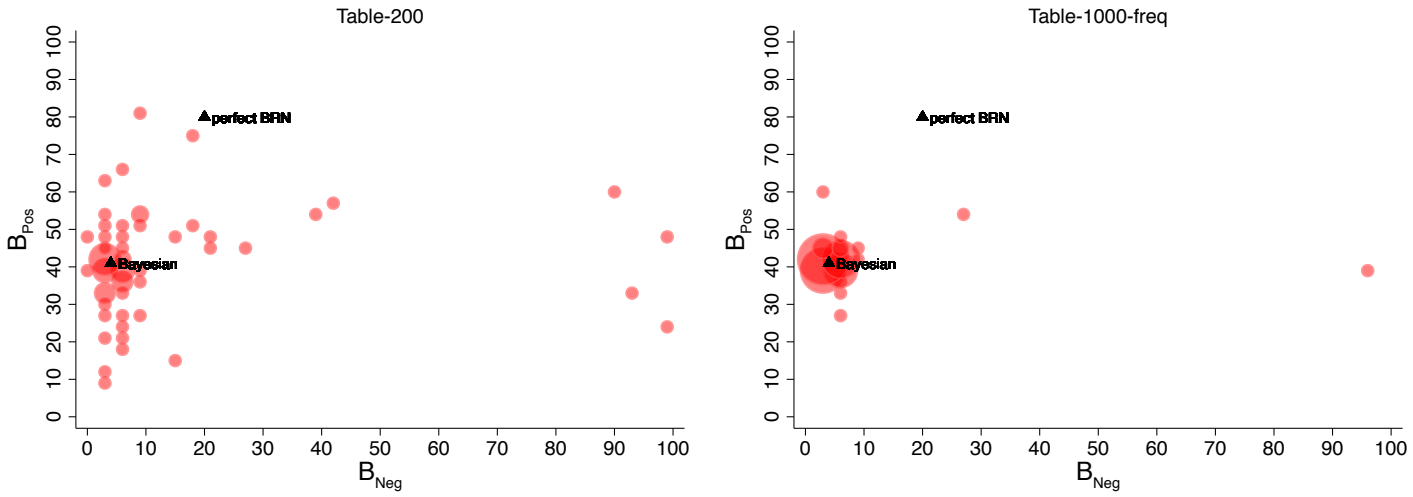


Figure 14: Density plots in the Primitives treatment

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.

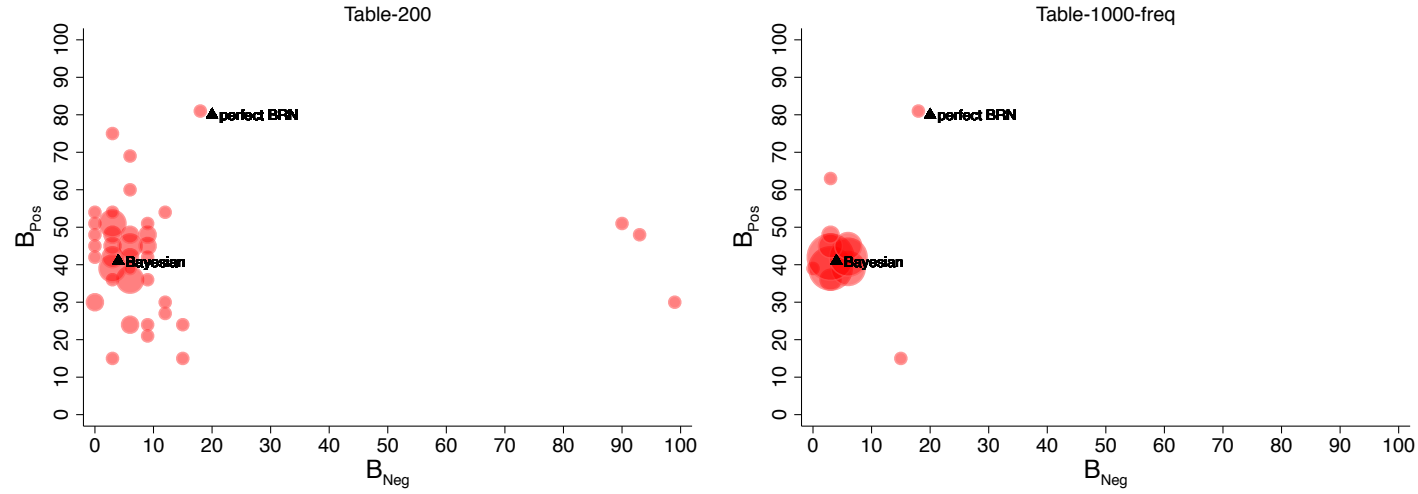


Figure 15: Density plots in the Primitives treatment

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 of recollection of feedback

	Positive	Negative	Total
Success	24	6	*30*
Failure	34	136	*170*
Total	*58*	*142*	*200*

(a) Actual average realizations

	Positive	Negative	Total
Success	36	17	*53*
Failure	31	116	*147*
Total	*67*	*133*	*200*

(b) Average report for R1 pBRN in P

	Positive	Negative	Total
Success	24	15	*39*
Failure	32	129	*161*
Total	*56*	*144*	*200*

(c) Average report for R1 Others in P

	Positive	Negative	Total
Success	29	14	*43*
Failure	33	124	*157*
Total	*62*	*138*	*200*

(d) Average report for all subjects in NoP

Table 13: Memory: Average actual realizations and average reports by treatment

Table 13 presents average actual realizations and average reports for each of the four questions and by group of subjects. Overall, on average, what subjects recall in terms of what they experienced is not too far from actual realizations, but there are some important differences between Round 1 pBRN subjects and other subjects. Such differences are not large, but appear to be directionally consistent with how the beliefs of these subjects differ from the beliefs of other subjects. For example, Round 1 pBRN subjects report on average having observed more successes than failures conditional on the signal being positive (Table 13b), while this is not the case for subjects classified as Round 1 Others in *Primitives* (Table 13c) or for subjects in *NoPrimitives* (Table 13d). In addition, Round 1 pBRN subjects report fewer failures conditional on the signal being negative relative to others.

Table 14 provides tests of statistical significance for these comparisons. The table reports two sets of regressions. In each regression of the first set, subjects' recollection of the frequency of each these outcomes is used as the dependent variable. In each regression of the second set, the difference between recalled frequency and the actual observed frequency is used as the dependent variable. The right-hand side includes a constant and a dummy for whether the subject is classified as Round 1 Others in *Primitives* and a dummy that takes value 1 if the subject participated in *NoPrimitives*. Focusing on reported frequency of (Pos, Succ) and (Neg, Fail) outcomes, there is evidence that Round 1 pBRN subjects are statistically different from other (using either set of regressions).

Table 13 suggests that Round 1 pBRN subjects' recollection of the feedback to be more

	(1) Recollection of feedback			(2) Recollection - Observed feedback		
	(Pos, Succ)	(Neg, Succ)	(Neg, Fail)	(Pos, Succ)	(Neg, Succ)	(Neg, Fail)
$D_{\text{Round 1 Others}}$	-12.5	-2.5	13.9	-11.2	-3.3	13.7
	(.005)	(.610)	(.093)	(.012)	(.506)	(.084)
$D_{\text{NoPrimitives}}$	-6.7	-3.2	8.8	-6.4	-2.9	7.1
	(.067)	(.430)	(.200)	(.081)	(.479)	(.280)
Constant	36	6.0	115.5	11.2	11.3	-19.4
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)

Table 14: Differences in recollection of feedback across treatments

Notes: The regressions in column (1) use as dependent variables the reports that subjects submitted for the number of outcomes that they recall the outcome to have been: positive and successful (Pos, Succ), negative and successful (Neg, Succ), and negative and a failure (Neg, Fail). Subjects also report the number of times they recall the outcome to have been positive and a failure, but since choices are forced to add up to 200, only three of the four answers are independent and this report is omitted in the regressions. Regressions in column (2) use as dependent variable the difference between recollection of feedback and what subjects actually observed in the corresponding category by round 200. The right-hand side of each regression includes a constant and two dummy variables, each taking value 1 only if the subject is in *Primitives* and classified as Round 1 Others ($D_{\text{Round 1 Others}}$), or is in *NoPrimitives* ($D_{\text{NoPrimitives}}$). Coefficient estimates are reported in the corresponding row and the p-values associated with the null hypothesis that the coefficient equals zero are reported between parentheses. Each set of regressions is run as a system of equations using the seemingly unrelated regressions procedure.

noisy (farther from what they have actually experienced) relative to others and to be distorted in the direction of their beliefs. However, it important to note that these differences are not very large. An interesting observation is that the conditional frequencies implied by what Round 1 pBRN subjects recall about the feedback is closer to the Bayesian benchmark than the beliefs held by these subjects. For example, on average, these subjects recall the frequency of success conditional on a positive signal to be 54 percent. While this is still substantially different from the Bayesian benchmark of 41 percent, it is much closer than the average belief of these subjects conditional on this signal in round 200, which is approximately 60 percent.

E Estimates from a Learning Model

The goal of this section is introduce a simple learning model that we estimate separately for the different groups of subjects in our data set. The model describes an agent updating beliefs (on B_{Neg} and B_{Pos}) with prior in the Beta distribution using outcomes from a Bernoulli process.[52]

An agent is asked to report B_{Neg} and B_{Pos}. The correct values are B^*_{Neg} and B^*_{Pos}. For

[52]We use the Beta distribution because if we update the Beta distribution using a Bernoulli process, we remain in the Beta distribution.

200 rounds, the agent receives feedback on the signal-state realization. That is, every round, conditional on the signal realization being negative (positive) the probability that the state is 'success' is B^*_{Neg} (B^*_{Pos}).

The agent has a prior on B_{Neg} and B_{Pos} that can be represented with the Beta distribution. (α_k, β_k) are the parameters that characterize the agent's prior on B_k for $k \in \{Neg, Pos\}$.

The parameter $\sigma \in [0, 1]$ describes the agent's attentiveness to the data in rounds 1-200.[53] That is, we assume that by round r, if there were S^r_k (F^r_k) rounds experienced up to that point in which the signal was $k \in \{Neg, Pos\}$ and the state was 'success' ('failure'), the agent's posterior on B_k for $k \in \{Neg, Pos\}$ is characterized by the Beta distribution with parameters $\alpha^r_k = \alpha_k + \sigma S^r_k$ and $\beta^r_k = \beta_k + \sigma F^r_k$. Note that $\sigma = 1$, gives us the Bayesian posterior with perfect recall for a Beta prior updated using outcomes from a Bernoulli process with success probability of B^*_k. We impose that $\sigma = 1$ when feedback from the first 200 rounds is presented in table form in Part 6.

We assume that the agent always reports expected value of B_k, which takes the following simple form with the Beta distribution:

$$\mathbb{E}(B_k|\alpha_k, \beta_k, \sigma, S^r_k, F^r_k) = \frac{\alpha_k + \sigma S^r_k}{\alpha_k + \beta_k + \sigma S^r_k + \sigma F^r_k}.$$

Focusing on three different groups in our data set (Round 1 pBRN Subjects in *Primitives*, Round 1 Others in *Primitives*, All Subjects in *NoPrimitives*) we estimate the five parameters, $(\sigma, \alpha_{Neg}, \beta_{Neg}, \alpha_{Pos}, \beta_{Pos})$, using least squares estimation. That is, we find the values that minimize the following:

$$\sum_i \sum_{k=Neg,Pos} \left(\sum_{r=1,200} \left(B^r_k - \mathbb{E}(B_k|\alpha_k, \beta_k, \sigma, S^r_k, F^r_k)\right)^2 + \left(B^T_k - \mathbb{E}(B_k|\alpha_k, \beta_k, 1, S^T_k, F^T_k)\right)^2 \right)$$

where the values with superscript T denote either the subjects' response or the feedback observed once feedback from 200 rounds is summarized in table form.[54]

The results for the estimation are such that for subjects whose initial beliefs are perfectly

[53] A natural interpretation for σ is that it captures the recall rate for feedback: each observation is remembered with probability $\sigma$and forgotten (hence not accounted for in updating) with probability $1 - \sigma$.

[54] Note that the summation is first over i, all subjects categorized in a group (Round 1 pBRN Subjects in *Primitives*, Round 1 Others in *Primitives*, All Subjects in *NoPrimitives*), then over the signal realization k, and finally over the first 200 rounds including Part 6 separately.

consistent with BRN in *Primitives* (Round 1 pBRN) we find that $(\sigma, \alpha_{Neg}, \beta_{Neg}, \alpha_{Pos}, \beta_{Pos})$ $= (0.17, 12.8, 45.8,\ 7.0, 3.3)$; for others in *Primitives* (Round 1 Others) the estimates are $(\sigma, \alpha_{Neg}, \beta_{Neg}, \alpha_{Pos}, \beta_{Pos}) = (0.98, 18.7, 30.1, 9.9, 6.9)$; and for those subjects in the treatment without primitives (*NoPrimitives*) we have $(\sigma, \alpha_{Neg}, \beta_{Neg}, \alpha_{Pos}, \beta_{Pos}) = (1.00, 5.9, 10.8,$ $3.4, 1.7)$. Figure 16 below plots the implied beliefs at different stages of the experiment. The dotted line presents the prior for B_{Neg} and B_{Pos} in round 1, the dashed line presents the expected posterior (calculated using expected feedback from 200 rounds) in round 200. Finally, the solid line presents the expected posterior after observing the feedback from the first 200 rounds in table form. The vertical lines present the Bayesian values conditional on each signal. The difference between the dotted and dashed line represents how much subjects are able to learn from feedback presented as outcomes from natural sampling in the first 200 rounds. The difference between the dashed and solid lines represents the inefficiency in learning resulting from partial attentiveness to feedback.

Figure 16 reveals some important patterns, reinforcing some of our findings in earlier analysis. First, we observe the prior to be strongest for those subjects categorized Round 1 pBRN in *Primitives*. In contrast, other subjects in this treatment as well as those subjects in *NoPrimitives* are estimated to have much more diffuse priors. This is one of the factors that slows down convergence to the Bayesian benchmark for the subjects categorized as Round 1 pBRN. Second, the estimates reveal Round 1 pBRN subjects in *Primitives* to be much less attentive to the feedback. The estimated value for σ is 0.17 for these subjects in contrast to 0.98 and 1.00 for others in *Primitives* and *NoPrimitives*, respectively. This provides insight on why presenting the feedback in table form is highly effective particularly for those subjects in *Primitives*.

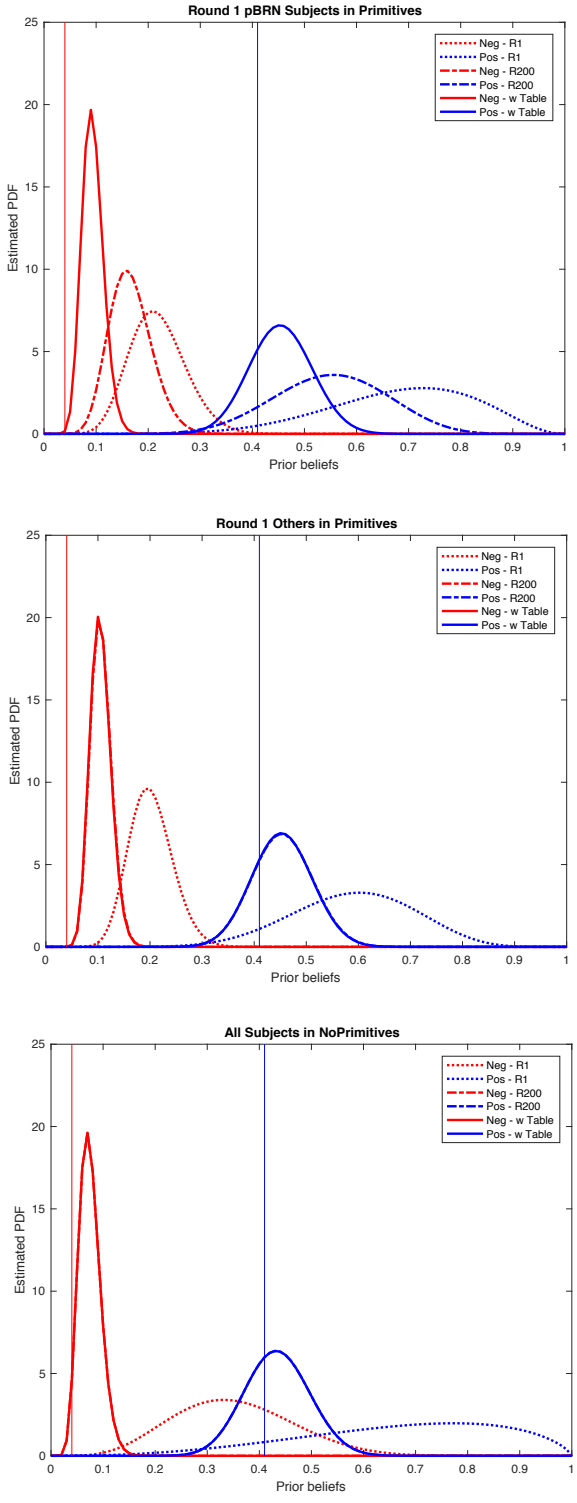


Figure 16: Estimation output

Notes: Each graph presents estimated beliefs at different stages of the experiment: round 1, 200, and the belief submitted after subjects observe the table summarizing feedback from rounds 1-200 (w table). The dotted line shows estimated prior for B_{Neg} and B_{Pos} (shown in red and blue, respectively). The dashed line shows posterior after at round 200: this is computed from the estimated prior using estimated attentiveness and expected outcomes from the first 200 rounds. The solid line shows posterior after after the subjects observe the summary table: this is computed from the estimated prior using expected outcomes from the first 200 rounds and setting attentiveness parameter to 1.