Online Appendix for

Mental Models and Learning:

The case of Base-Rate Neglect

Ignacio Esponda Emanuel Vespa Sevgi Yuksel

CONTENTS:

A. Literature review: Further details

I.A. Literature on base-rate neglect with feedback

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

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

Grether (1980, 1992) and Griffin and Tversky (1992) are some of the early economics-style experiments on the topic where subjects are financially incentivized to form accurate beliefs and the updating problems are presented in the standard framework of judging the likelihood of abstract events (for example, event involving balls drawn from different urns). Importantly, Grether (1980) also introduces a general way of measuring partial base-rate neglect based on regression analysis focusing on the log likelihood ratio of different events. This approach is now commonly used in many papers, including this one, studying updating behavior. It should be noted that none of these early papers studied how behavior changes with feedback. In most experiments subjects only answered one belief updating question, and in others that included multiple questions, the parameters and/or

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

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

the environment changed between questions with no feedback between questions.

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

Third, more closely related to our research question, Section 2.1.2 of Koehler (1996) discusses several early experiments where subjects have an opportunity to learn about base rates from direct feedback. For example, Manis et al. (1980), Lindeman et al. (1988), and Medin and Edelson (1988) provide evidence that base rates influence probabilistic judgements more when they are directly experienced through trial-by-trial outcome feedback. None of these papers include a treatment that can be mapped back cleanly to either of our treatments, but they provide insights that parallel some of our findings. In Manis et al. (1980) subjects were shown 50 yearbook pictures of male students and, for each randomly selected picture, they were asked to predict the person's position on two issues (marijuana legalization and mandatory seatbelt legislation). Note that a signal in this context can be interpreted to be the characteristics of person observed in the picture. The informativeness of these pictures is ambiguous and actually manipulated to be non-existent. The results suggest that subjects adjust their judgments in response to the accuracy of their past predictions. In Lindeman et al. (1988) subjects are given 16 different versions of Kahneman and Tversky's engineer-lawyer problem. While the analysis indicates that feedback leads to adjusted probability estimates closer to the Bayesian benchmark, the type of feedback that subjects are provided is highly unnatural and unusual.[61] It is also important to note that the paper does not find any transfer of learning in this environment to another one where subjects can display base-rate neglect (based on Zukier and Pepitone 1984). Medin and Edelson (1988) report results from an experiment where the task involved participants diagnosing hypothetical diseases on the basis of symptom information. It is difficult to interpret their results as their learning environment is complicated by the fact that there are many features of the environment that are varied within subjects and some of these involve ambiguous signals. Overall, they find mixed results for subjects incorporating the base rate. Among these set of papers, the closes to our work is Christensen-Szalanski and Beach (1982). The paper demonstrates that subjects make use of base rates in forming posterior probabilities when they have experienced the relationship between the base rate and the diagnostic information, but fail to make use of the base rate when they only experience the base rate and are given the reliability of the signal.[62]

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

[62] One of their treatments (where subjects experience both the state and the signal in direct feedback) is similar to our *NoPrimitives* treatment where subject are not given the primitives and learn from feedback. However, a critical

Since the review article of Koehler (1996), there has been a considerable literature in psychology studying whether subjects can learn through direct experience to incorporate base rates into posterior beliefs. These papers are reviewed in Goodie and Fantino (1999). While this body of work often provides evidence that subjects can learn from experience to adjust actions towards optimal behavior, the approach in these papers are fundamentally different from ours. The framework adopted in most of these experiments is one where subjects repeatedly choose between two binary options after observing a binary cue, receiving feedback about the optimality of the choice after each round. The choices are often between abstract options (for example, green or blue) and the cues could be labeled similarly or differently from the options (for example, matching colors or arbitrary shapes). Critically, subjects are not informed about the primitives determining statistical relationship between the cue and the optimal action.[63] In this respect, these experiments are closest to our *NoPrimitives* treatment in which the prior and the reliability of the signal were not provided to the subjects. However, there are still some differences in how such a treatment is implemented in these papers that could be important for behavior. For example, in these experiments, subjects are not told explicitly that the environment they face repeatedly is a stationary one in the sense that each round corresponds to an independent draw of optimal action/cue pair from the same distribution. Note also that the learning problem is different from the one we study in that in these experiments subjects can possibly learn the optimal binary action conditional on each signal without ever forming precise beliefs conditional on each signal.

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

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.

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

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

Literature on probability matching & feedback

The experimental literature on probability matching is surveyed in Vulkan (2000) and, more recently in Erev and Haruvy (2013). Most papers in the early literature on probability matching used an environment in which the primitives were not provided to subjects. To illustrate, here is a typical example taken from Erev and Haruvy (2013). There is an event E that happens with probability 0.7, but subjects do not know this probability. In a given round, subjects click on button H or button L. Button H pays (L pays) a positive amount if E occurs (E does not occur). After 50 rounds, the observed rate of H selection is 70%. This finding coincides with the earlier literature in which subjects were reported to make choices that are close to 'probability matching' instead of optimizing. However, more recent papers have demonstrated that longer experience slowly moves choices toward maximization. In that example, the H rate in the data was 90% between rounds 51 and 150. These findings are consistent with our long-run findings for the *NoPrimitives* treatment.

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

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

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

There is a large set of experiments in which feedback of some sort plays a role but where feedback is not part of the central object of study. Meanwhile, the literature that focuses specifically on feedback can perhaps be organized into two groups. The first is the relatively large experimental literature that studies how people use feedback to learn in games, which dates at least back to Harrison and Hirshleifer (1989) and Prasnikar and Roth (1992), and is less directly related to our work in this paper. Models such as reinforcement learning (Erev and Roth (1998), Roth and Erev (1995)), directional learning (Selten and Stoecker 1986), adaptive learning (Cheung and Friedman 1997), experience-weighted attraction (Camerer and Hua Ho 1999), and rules-based learning (Stahl 2000) were proposed and tested in this literature. The focus is on what kind of model can rationalize how people learn from feedback, mostly in settings in which taking into account the behavior of other players is crucial. For a detailed survey of the literature, see Part 5 of Dhami (2020).

The second group involves a more recent set of papers that are closer to our paper and focus

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

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

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

Cognitive limitations of agents are behind the second case capturing long-run play that deviates from Nash play. For example, the notion of Behavioral Equilibrium (Esponda 2008) captures the long-run behavior of an agent that has difficulties to understand endogenous selection in her feedback. An experimental test of these predictions is studied in Esponda & Vespa (2018). An agent who does not control for selection will have a biased view of the environment. Such biased view would lead to decisions that are suboptimal, but could generate feedback that results in a Non-Nash equilibrium. The experimental test consists of comparing choices in a treatment in which feedback involves a selected sample against a treatment in which such selection is not present. The evidence suggests that most subjects do not adjust for selection and end up making suboptimal choices in the long run.

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

Relatedly, Barron et al. (2019) study a situation in which individuals try to learn from observing behavior of others who have faced similar decisions previously. However, information from others involves selection because choices of others are observed conditional on private information. Their experimental paper uses the theoretical selection neglect framework of Jehiel (2018). The paper documents evidence of selection neglect, which is consistent with findings in other papers in this

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

literature. They also document that issues with selections increase when the agents generating the feedback that others use have more private information.[67] In all of the papers in this part of the literature the quality of the feedback depends on subject's choices. A difference with our paper is that in the environments we study the quality of subjects' choices is independent of the quality of the feedback that subjects receive.

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

B. Details on the experimental design

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

Core treatments

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

Part 0

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

Part 1

The aim of this part is to introduce the strategy method. There are two decks of cards, each with 100 cards and cards can be green or blue. One card of the 200 cards is randomly selected and they have to indicate the chance that the selected card is green vs. blue in case it belongs to deck 1, and separately, in case it belongs to deck 2. On the screens subjects are informed of the composition of each deck before they submit their answers. As the problem in Part 0, there is an objective answer to maximize payoffs in this problem. After they submit their answers, an explanation appears on the screen describing the answers that maximize payoffs. They repeat this problem twice, each time with different compositions of each deck.

Part 2

This section involves the main task. For each possible test result (positive, negative) participants submit the chance that the project is a success vs. a failure. The instructions are presented in

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

Appendix J. This is the only part in the experiment where the instructions to treatment *Primitives* differ from those of *NoPrimitives*.

Part 3

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

Part 4

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

Part 5

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

Part 6

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

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

Part 7

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

Part 8

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

Part 9

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

Survey

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

Mechanism treatments

Primitives w/ shock

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

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

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

These treatments are identical to the Primitives and *NoPrimitives* treatment, respectively until Part 3. At that point and for both treatments, the instructions for Part 3 include the following sentences in the last paragraph: "(...) You will also have a 'lock-in' option. This option enables you to use your current choices for the current round and all future rounds. In other words, if you select this option, you will not need to click through all the remaining rounds; instead you will jump to the end of the experiment. But this also means that you will not be able to modify your choice for future rounds. Note that even if you use the 'lock-in' option to skip to the end of

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

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

Table 2: Summary of BRN Treatments

the experiment, you will not be able to leave early. We will pay you only after everybody is done. You will be able to make choices at your own pace in this part. Part 3 will end after you make your choices for all rounds." Given the option to lock in choices, we merged parts 3 and 4 in this treatment. Essentially, subjects were told that in Part 3 they would face additional 199 rounds.

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

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

Voting treatments

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

Part 1

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

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

Part 2

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

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

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

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

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

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

Table 3: Summary of Voting Treatments: Part 1

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

III.A. Treatment differences in rounds 1-200

Statistical analysis on treatment differences

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

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

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

Aggregate measure of partial base-rate neglect

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

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

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

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

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

estimate on α gives us a continuous measure of the level of partial base rate neglect in the aggregate data.[71]

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

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

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

Figure 11a separately follows with diamonds the behavior of Round 1 pBRN subjects. Note that, by definition, all Round 1 pBRN subjects make pBRN choices in round one, so that the starting point for this group is $(B_{Pos}, B_{Neg}) = (80, 20)$. While beliefs for these subjects move towards the Bayesian benchmark with experience, by round 200 beliefs for these subjects are substantially farther away from the Bayesian benchmark relative to the average in *Primitives*. Furthermore, the beliefs of Round 1 pBRN subjects are significantly different from subjects in *NoPrimitives*. This is shown in column (1) of Table 7; for example, there is a significant fifteen percentage-point difference

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

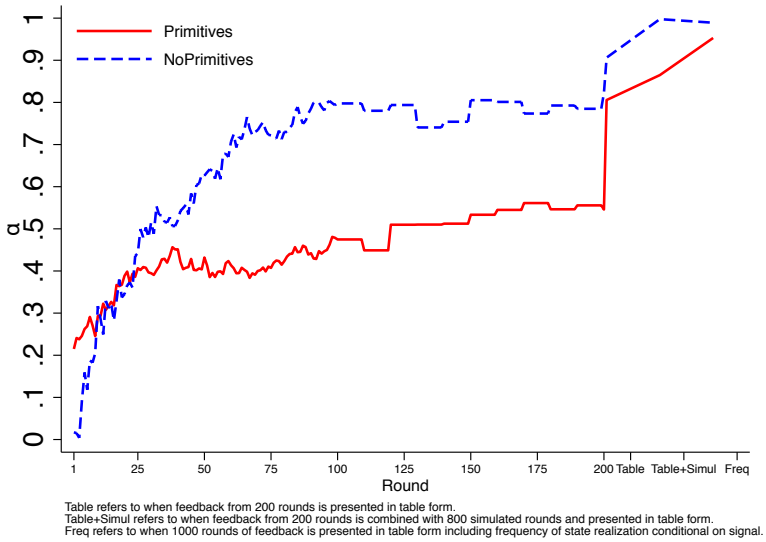
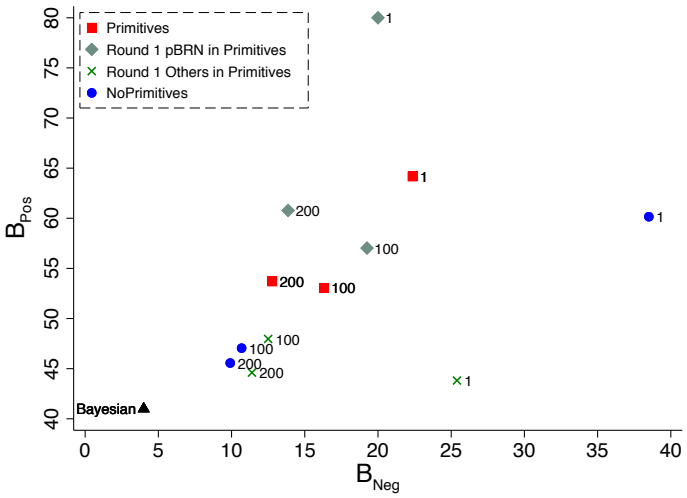


Figure 10: Estimates of α per round by treatment

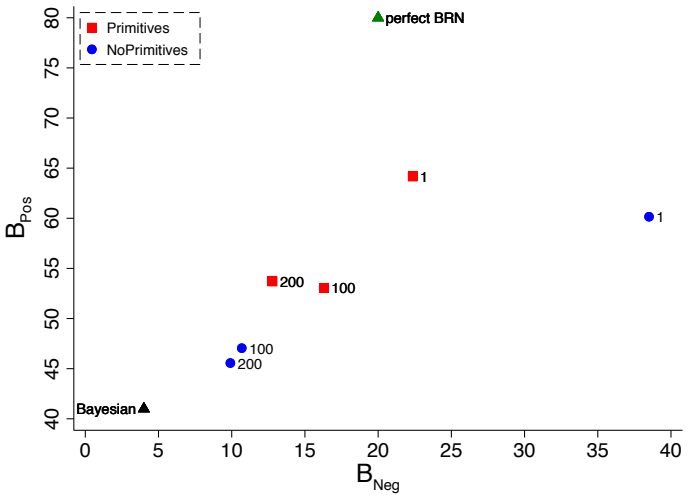
	Estimates						
	P	*NP*	*Ps*	*Pl*	*NPl*	*Pf*	*NPf*
α	0.55	0.82	0.82	0.49	0.72	0.89	0.99
β	0.87	0.88	0.81	0.77	0.84	0.83	0.99
	Differences						
	P vs. *NP*	*P* vs. *Ps*	*NP* vs. *Ps*	*Pl* vs. *NPl*	*Pf* vs. *NPf*	*Pf* vs. *P*	*NPf* vs. *NP*
α	$p = 0.001$	$p = 0.001$	$p = 0.987$	$p = 0.014$	$p = 0.127$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

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

Table 6: Estimates from Grether Regressions in Round 200



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



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

Figure 11: Evolution of submitted beliefs by subgroups

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Triangles indicate the Bayesian and the pBRN benchmarks. Squares (Circles) report averages in treatment *Primitives* (*No-Primitives*). 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.

in the average of B_{Pos} between the two groups.[72] Here, we focus on Round 1 pBRN subjects who made pBRN choices in round 1, but may change their behavior as the session evolves. Additionally, it is possible to trace the proportion of subjects in each round who make choices consistent with pBRN. Such evolution is presented in Figure 12.

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

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

Table 7: Estimation output for subsets of subjects

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

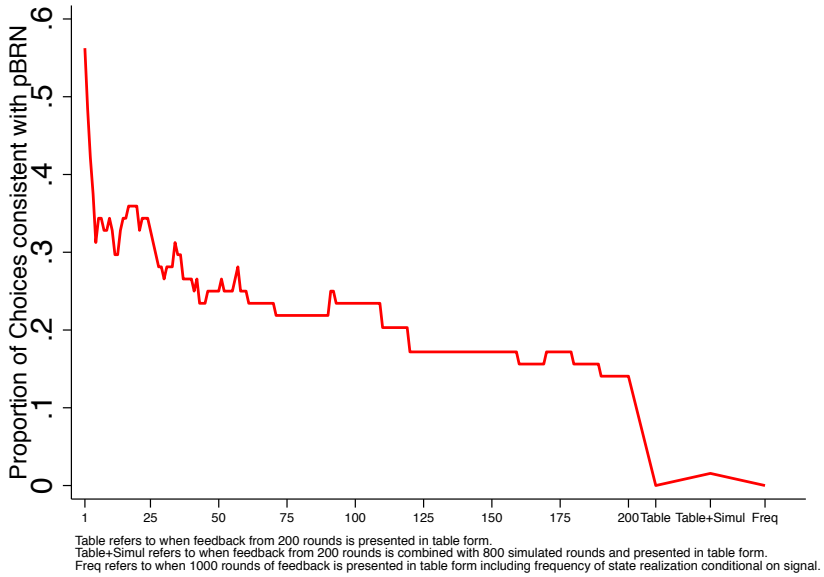


Figure 12: Proportion of choices consistent with pBRN in *Primitives* as the session evolves

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

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

Table 8: Estimation output for subsets of subjects

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

labels) satisfy the constraint. Even within this subset, large treatment differences emerge by round 100, and these differences remain by round 200. Table 8 verifies these patterns statistically.

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

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

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

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

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

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

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

Table 9: Recollection of feedback

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

Convergence and time

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

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

III.B. Treatment differences after round 200

Recollection of feedback

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

A first look at results is presented in Table 9a, which shows the average implied frequency of

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

success conditional on each signal calculated from subjects' recollection of feedback and, in the first row, the actual average frequencies that subjects observed.

We find that frequencies implied by the recollection of feedback are farthest away from the actual frequencies for Round 1 pBRN subjects. Note also that for these subjects the frequencies implied by the recollection of feedback deviate from actual frequencies precisely in the direction of the beliefs they submit.[75]

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

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

Summary tables

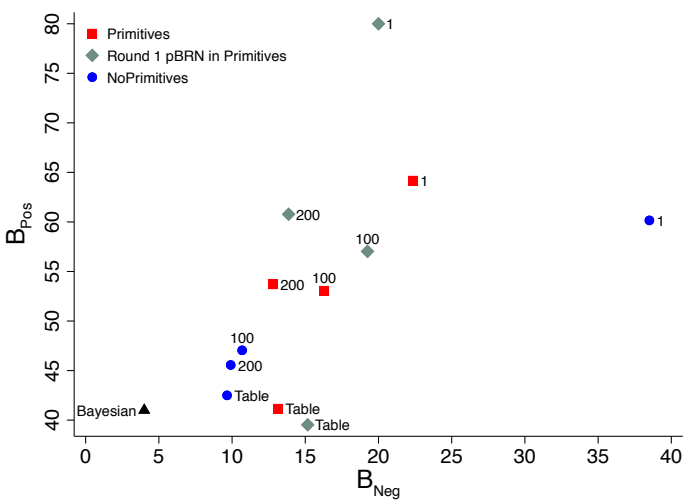
In this section we study the effect of showing subjects aggregate data (that they have already experienced) in a summarized table form. As explained in Online Appendix B, we begin by presenting subjects with feedback from rounds 1-200 using a two-by-two table that reports the number of rounds that each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

The main finding is that introducing the table dramatically moves beliefs closer to the Bayesian benchmark in *Primitives*, particularly with respect to B_{Pos}. The movement of average beliefs can be observed in Figure 13a, in which the average belief for this part of the experiment (denoted 'Table') is shown for different groups. While there is no significant change with respect to B_{Neg}, we observe a downwards adjustment in B_{Pos} of approximately 14 percentage points in treatment *Primitives*.

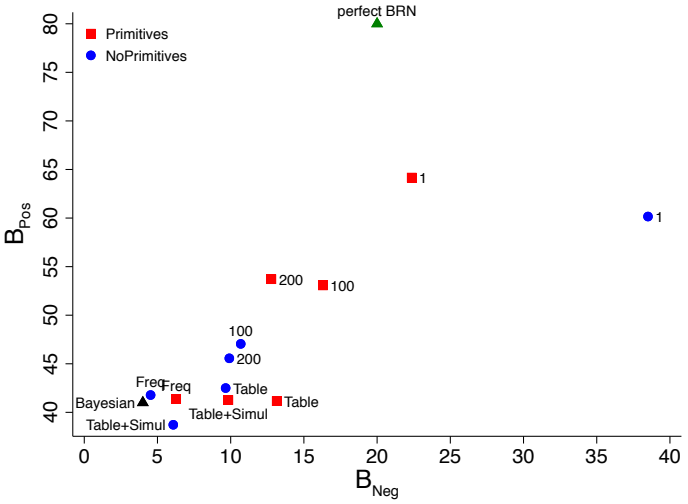
[75] This is consistent with subjects using their mental model (due to their limited recollection of past events) to reconstruct what might have happened to them.

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

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



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



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

Figure 13: Reported beliefs at different parts of the session

Notes: The vertical (horizontal) axis represents beliefs conditional on the signal being positive (negative). Triangles indicate the Bayesian and pBRN benchmarks. Squares (Circles) report averages in *Primitives* (*NoPrimitives*). The numbers indicate the round for which the averages are reported. 'Table' refers to when subjects are presented with a summary table of the feedback collected in 200 rounds. 'Table + Simul' refers to when the summary table includes 800 additional simulated rounds (for a total of 1000 rounds). 'Freq' refers to when subjects see the table with 1000 rounds of feedback and the relevant frequencies.

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

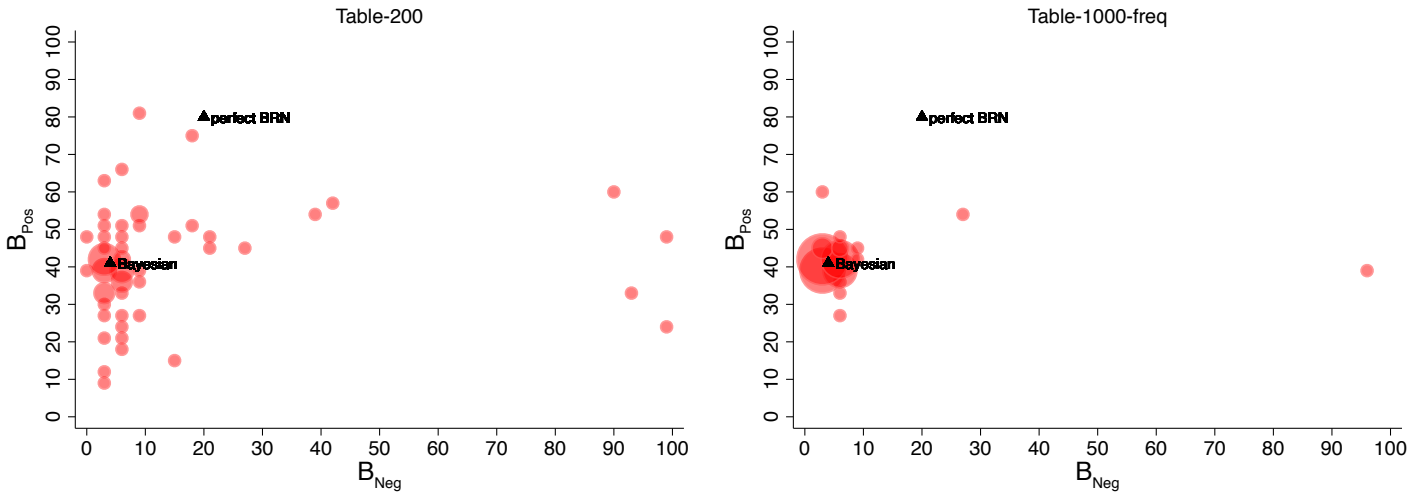


Figure 14: Density plots in the Primitives treatment

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Beliefs are presented at the individual level (rounded to multiples of 3). The size of each bubble represents the number of subjects with such beliefs. Table-200 refers to when feedback from 200 rounds is presented in table form. Table-1000-freq refers to when 1000 rounds of feedback is presented in table form including frequency of state realization conditional on signal.

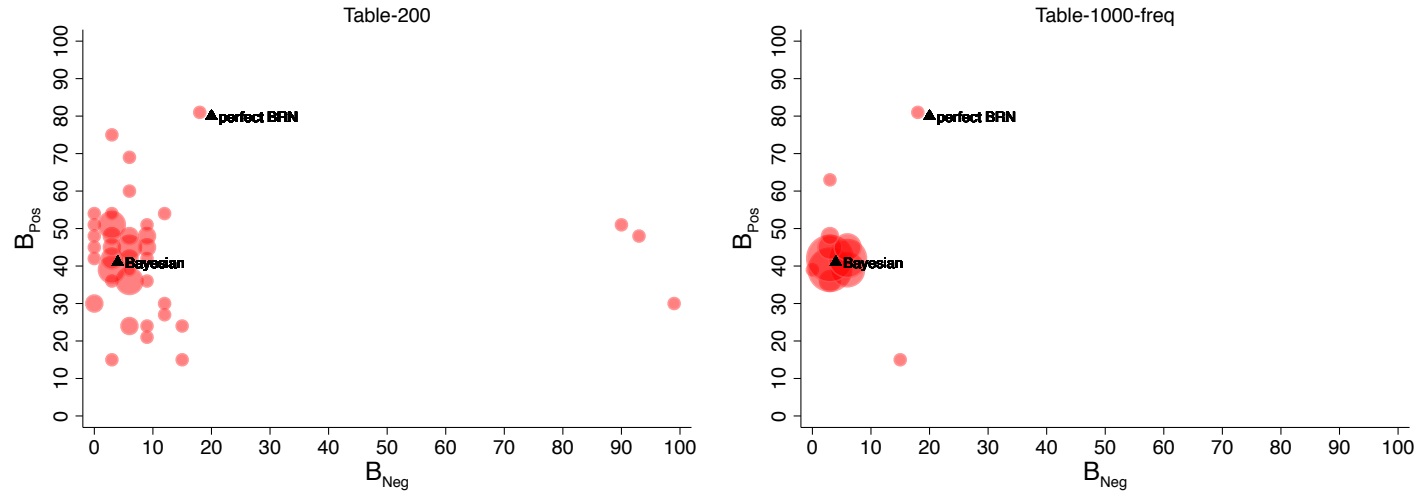


Figure 15: Density plots in the Primitives treatment

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Beliefs are presented at the individual level (rounded to multiples of 3). The size of each bubble represents the number of subjects with such beliefs. Table-200 refers to when feedback from 200 rounds is presented in table form. Table-1000-freq refers to when 1000 rounds of feedback is presented in table form including frequency of state realization conditional on signal.

D. Additional analysis: Confidence

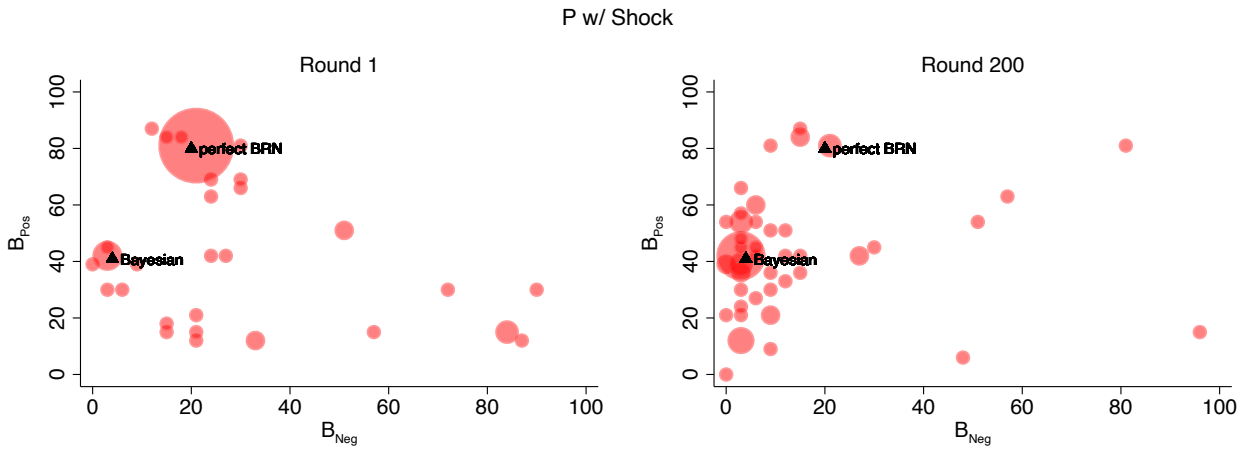


Figure 16: Density Plots for *Primitives w shock*

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

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

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

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

E. Additional analysis: Attentiveness

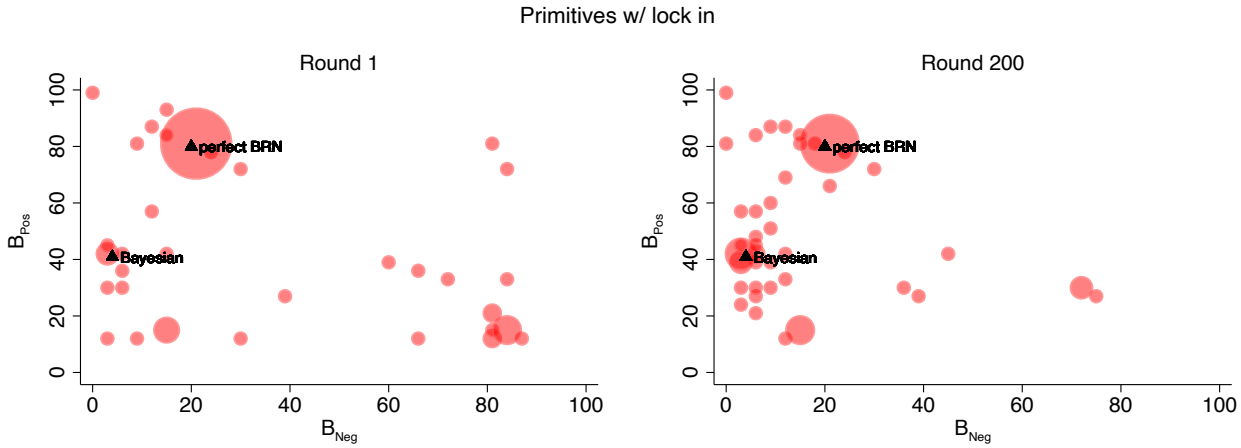


Figure 17: Density Plots for *Primitives w/ lock in*

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

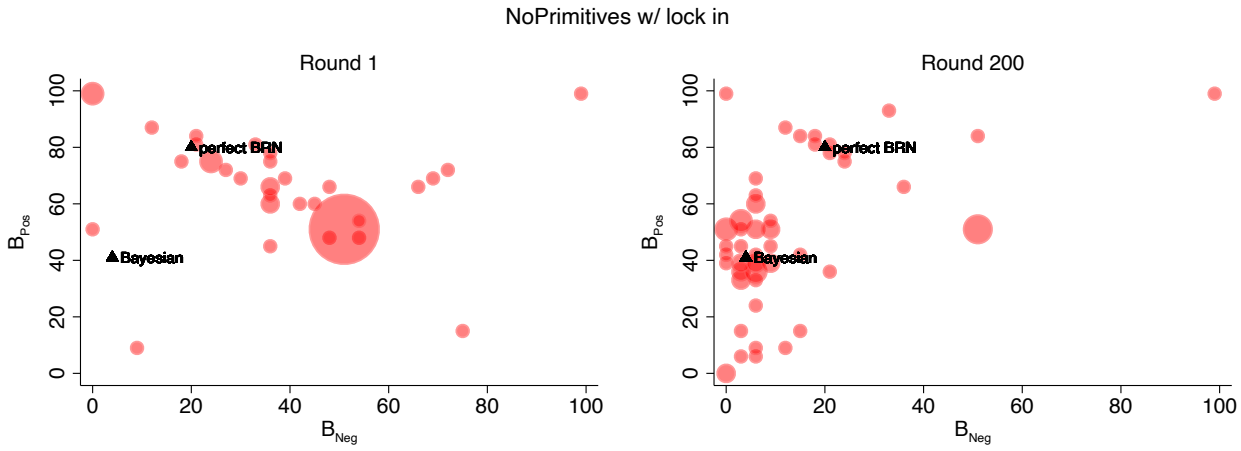


Figure 18: Density Plots for *NoPrimitives w/ lock in*

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

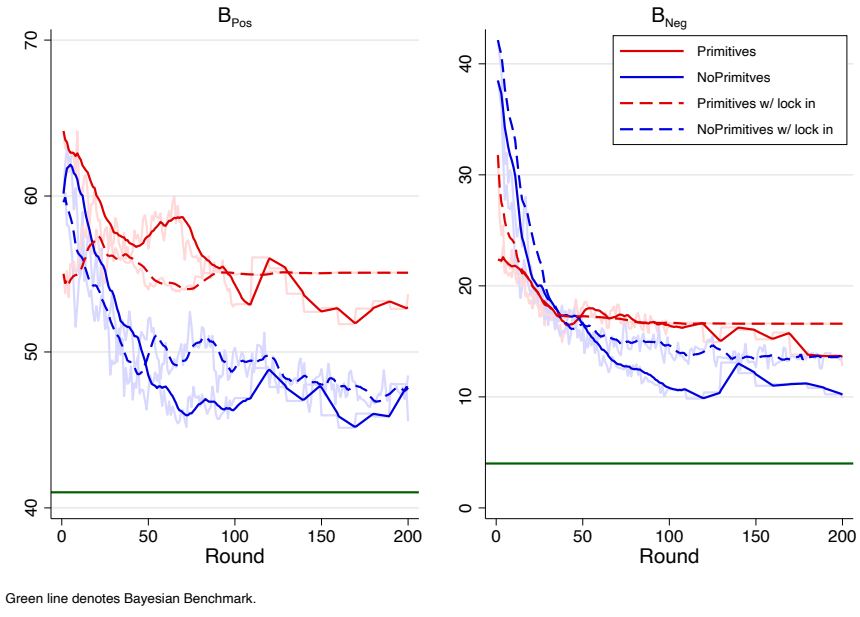


Figure 19: Evolution of Beliefs in Treatments with Lock In

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

F. Additional analysis: Costly attention

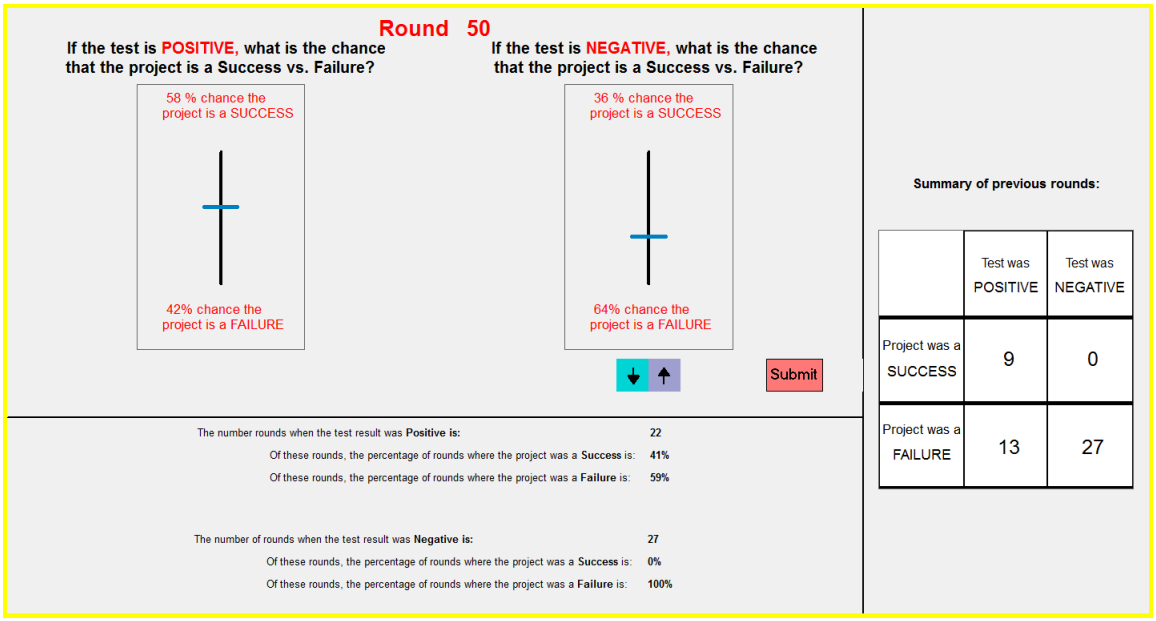


Figure 20: Interface Screenshot of Treatments with Frequencies (Round 50)

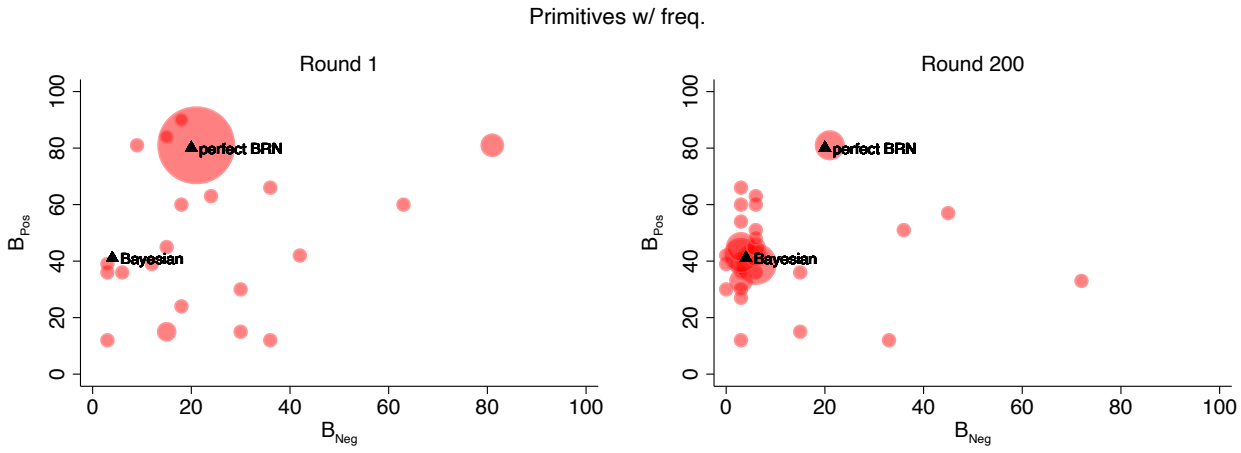


Figure 21: Density Plots for *Primitives w/ freq*

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

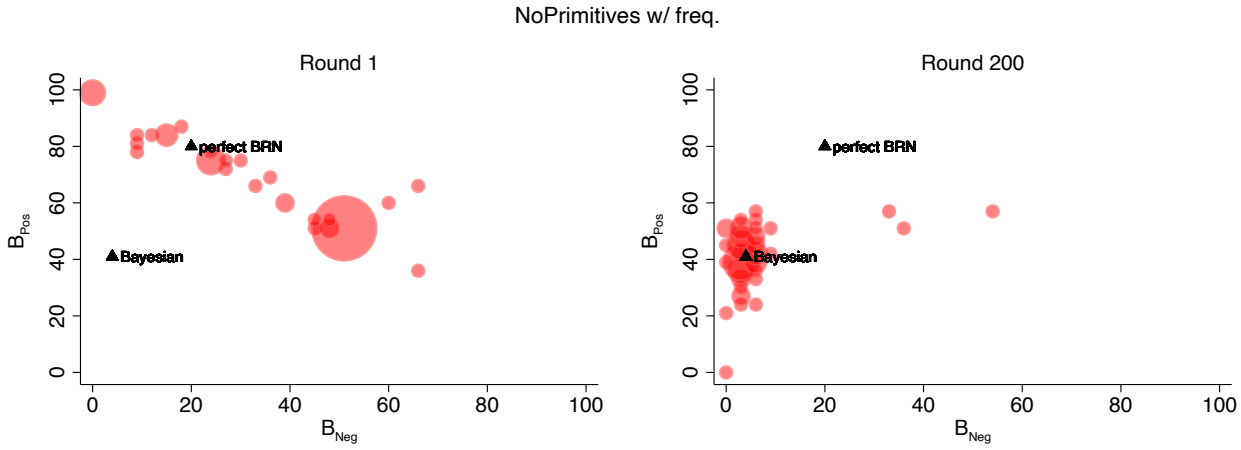


Figure 22: Density Plots for *NoPrimitives w/ freq*

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

Details on estimation of the learning model on the aggregate level

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

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

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

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

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

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

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

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

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

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

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

Estimates on long-run outcomes

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

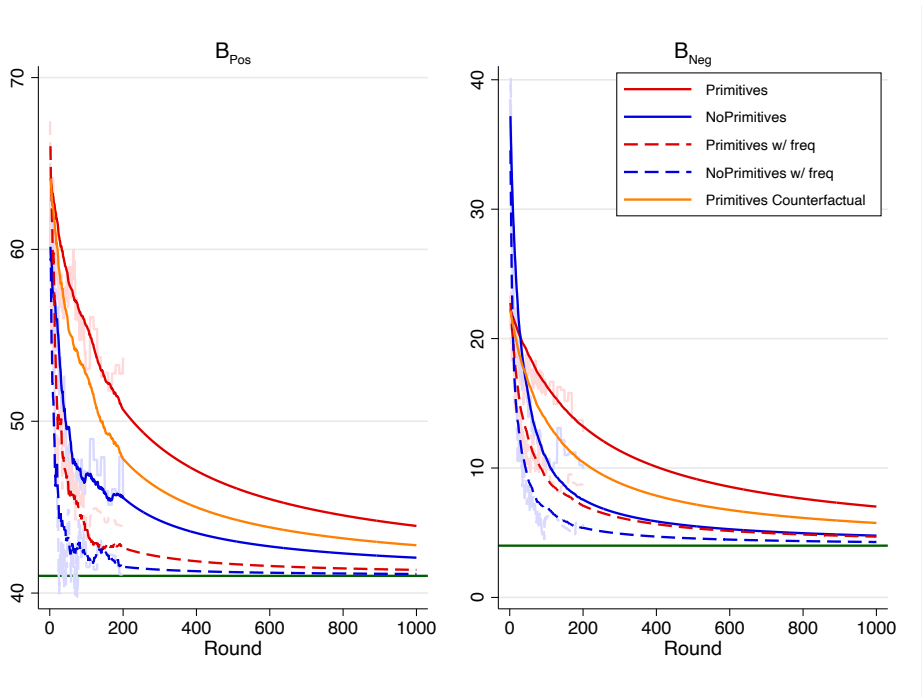


Figure 23: Model Estimates on Evolution of Beliefs for Rounds 1-1000

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

Estimating the learning model on the individual level

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

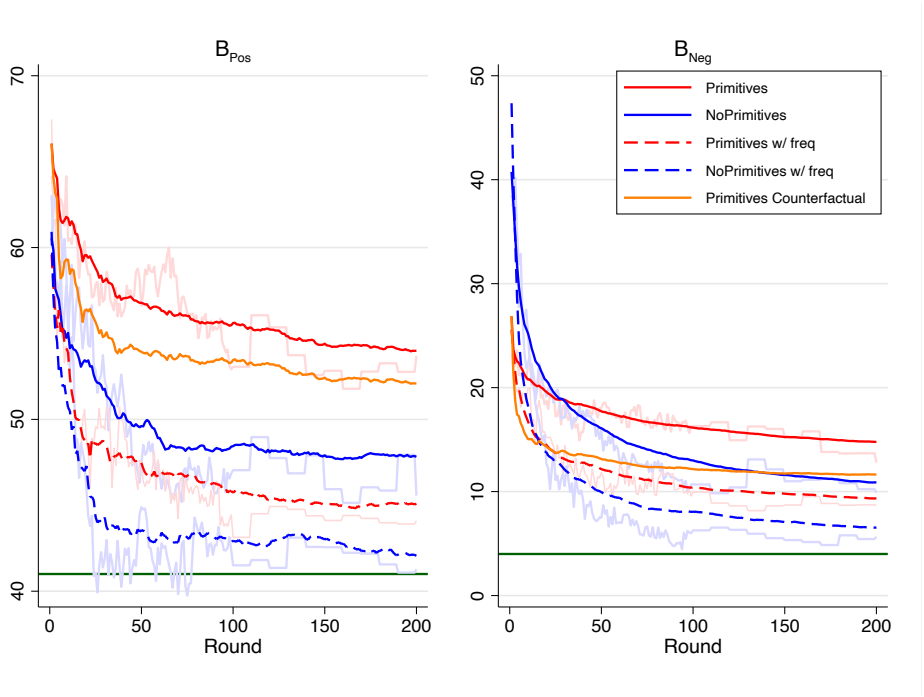


Figure 24: Model Estimates on Evolution of Beliefs Accounting for Heterogeneity

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

G. Additional analysis: Heterogeneity

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

Here we summarize key findings:

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

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

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

4. In treatments with lock-in option, when primitives are provided, those who start at the BRN point lock-in slightly later than others ($p = 0.079$) in the same treatment, but much earlier than those who are not given primitives ($p < 0.001$). However, subjects who start at the BRN point do not keep revising their beliefs for significantly longer than others in the same treatment (but both groups stop revising earlier than those who are not given primitives).[78] This indicates that information on primitives lowers engagement with the data for *all* subjects (both those who start at the BRN point and others). This suggests that subjects classified as others in treatments with primitives learn both from data and from primitives. See Tables 15 and 16 for details.

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

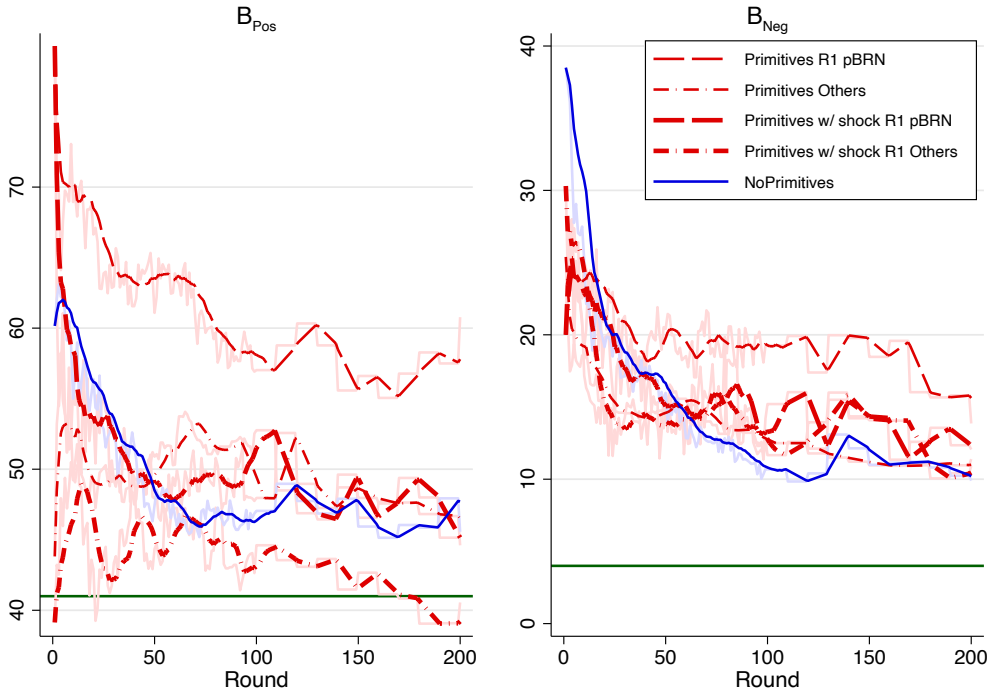


Figure 25: Evolution of Beliefs for R1 pBRN Subjects and Others in *Primitives* and*Primitives w/ shock* vs. *NoPrimitives*

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

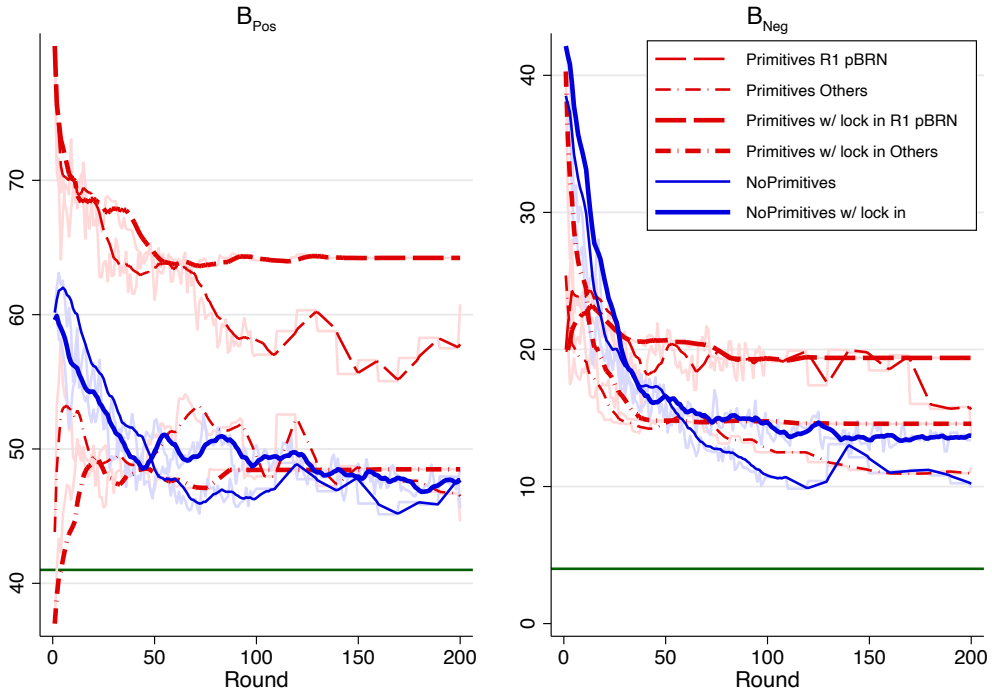


Figure 26: Evolution of Beliefs for R1 pBRN Subjects and Others in *Primitives* and *Primitives w/ lockin* vs. *NoPrimitives*

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

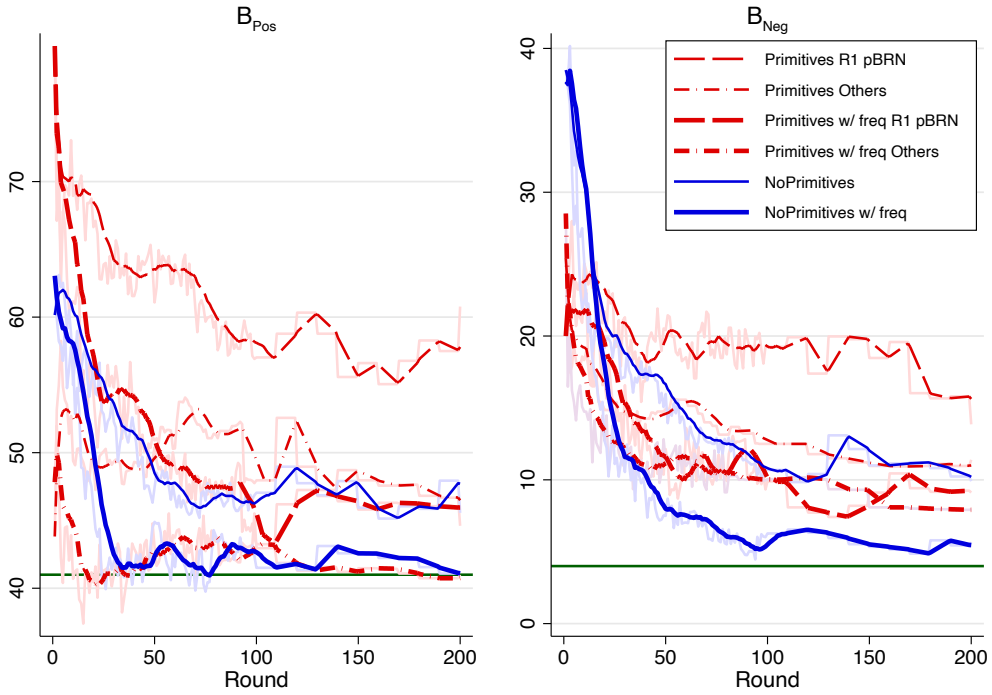


Figure 27: Evolution of Beliefs in *Primitives*, *Primitives w/ freq*, *NoPrimitives* and *NoPrimitives w/ freq*

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

Table 12: Round 50

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

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

Table 13: Round 100

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

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

Table 14: Round 200

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

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

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

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

Standard errors in parentheses.

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

(1): Data from Primitives and NoPrimitives.

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

(3): Data from Primitives w/ shock.

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

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

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

Standard errors in parentheses.

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

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

H. Additional analysis: Transfer learning

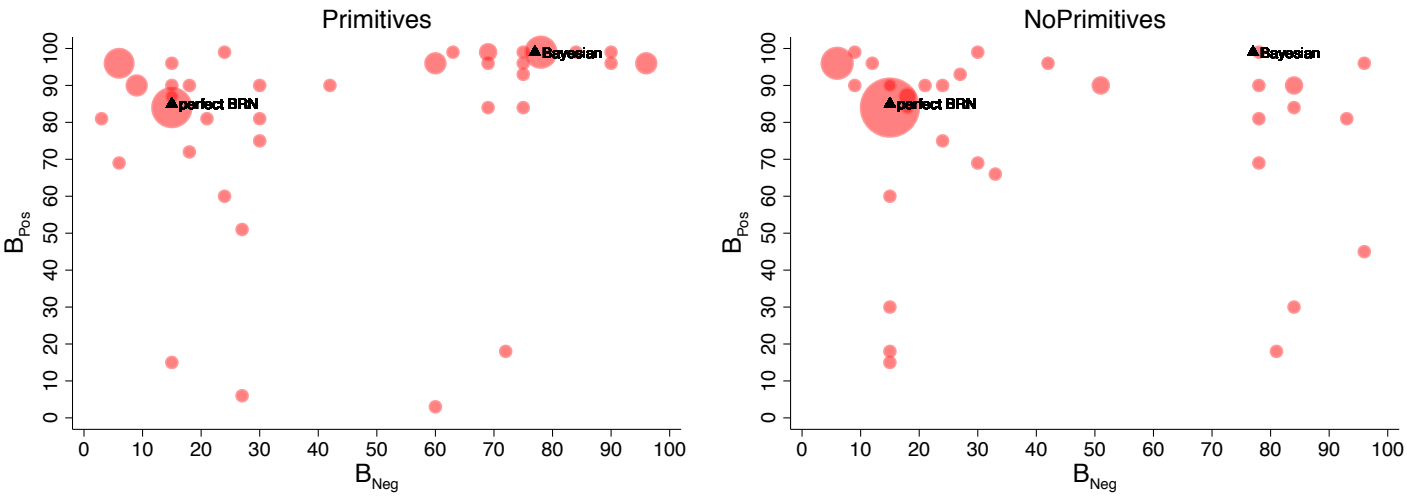


Figure 28: Transfer Learning: Density Plots in Final Round with New Primitives

Notes: The vertical (horizontal) axis captures the belief conditional on the signal being positive (negative). Beliefs are presented at the individual level (rounded to multiples of 3). The size of each bubble represents the number of subjects with such beliefs. The data is from the final round of the core treatments where the prior and the reliability of the signal is changed.

I. Additional Analysis: Evidence beyond the updating problem

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

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

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

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

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

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

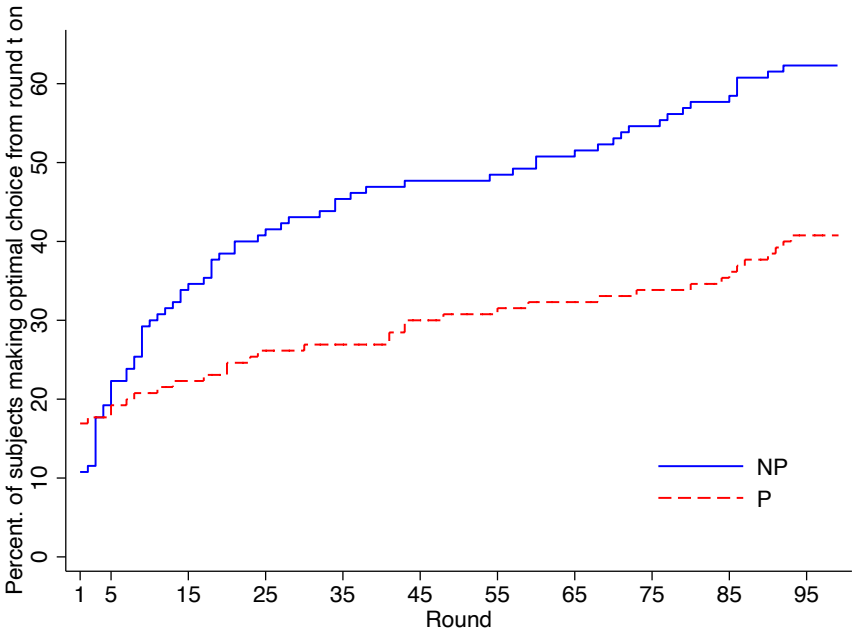


Figure 29: Subjects making optimal choices in *Primitives (Voting)* and *NoPrimitives (Voting)*

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

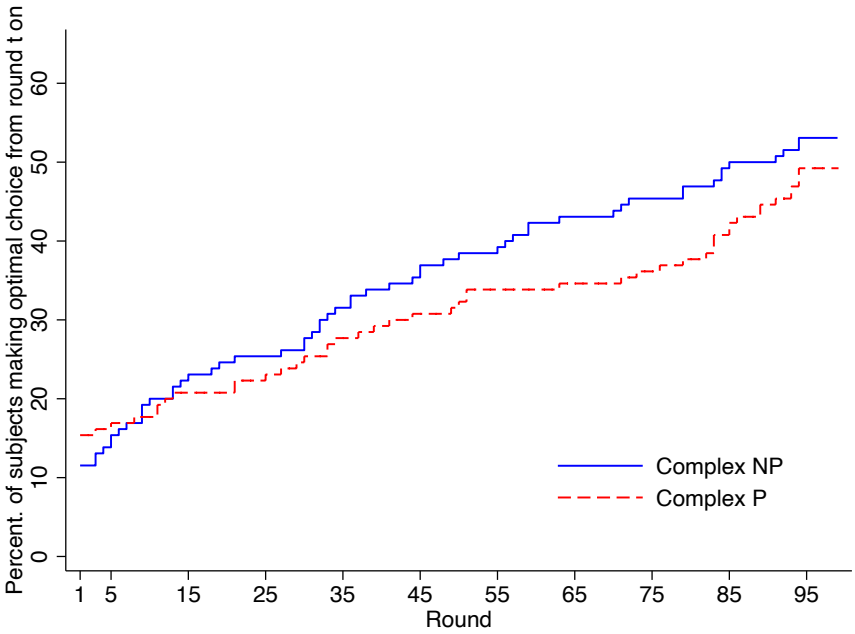


Figure 30: Subjects making optimal choices in *Complex Primitives (Voting)* and *Complex NoPrimitives (Voting)*

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

J. Experimental instructions

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

Round 1 Instructions:

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

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

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

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

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

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

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

That means that:

[*Primitives*:

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

[*NoPrimitives*:

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

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

We will ask you to submit two assessments:

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

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

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

Round 1 screenshot (part 2 in instructions):

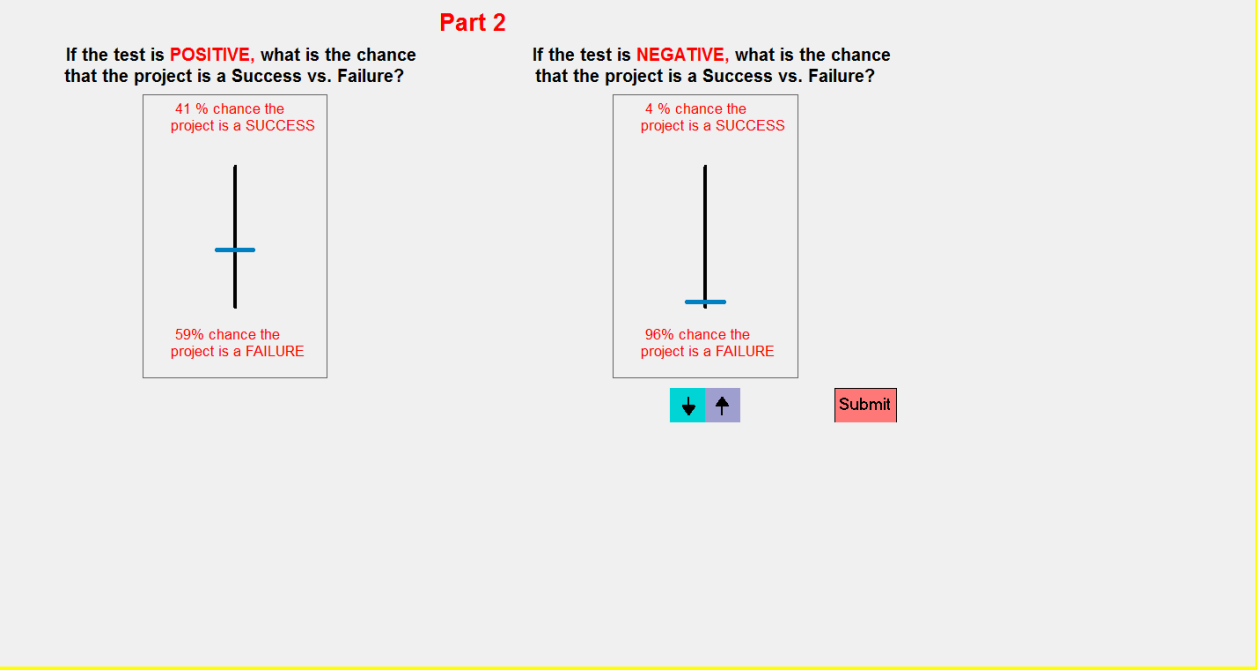


Figure 31: Interface screenshots for round 1 (presented as part 2 to subjects)

Appendix References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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