Disentangling Suboptimal Updating: Task Difficulty, Structure, and Sequencing.*

Marina Agranov† Pëllumb Reshidi ‡

June 7, 2024

Abstract

We study underlying reasons for the failure of individuals to adhere to Bayes' rule and decompose this departure into three elements: (i) task difficulty, (ii) information structure, and (iii) timing of information release. In a series of controlled experiments, we systematically alter all three elements and quantify their magnitude. We link task difficulty with the degree of non-linearity embedded in Bayesian updating. We experimentally explore this link and find empirical support for it.

*Agranov gratefully acknowledges the support of NSF grant SES-2214040. We thank Larbi Alaoui, Ben Enke, Alessandro Lizzeri, Kristof Madarasz, Kirby Nielsen, Pietro Ortoleva, Leeat Yariv, and participants at many seminars and conferences for helpful comments and suggestions.

†Department of Economics, California Institute of Technology and NBER; magranov@hss.caltech.edu

‡Department of Economics, Duke University pellumb.reshidi@duke.edu

1 Introduction

Across various contexts, updating beliefs is critical in the decision-making process. In the last few decades, economists and psychologists have made significant progress in understanding how people engage with new information. In many situations, the standard Bayesian updating framework accurately describes the evolution of beliefs (Grether, 1978; Camerer, 1987; Charness and Levine, 2005).[1] At the same time, there are frequent and consistent deviations from Bayesian updating. These deviations persist even when people have ample opportunities to learn (Esponda et al., 2023) and when stakes are very high (Gneezy et al., 2023). Moreover, these deviations are prevalent even among professionals who frequently deal with such issues (Benjamin, 2019). What makes some belief-updating tasks more difficult and more mistake-prone than others? This is the focus of our paper.

We present findings from a series of lab experiments involving belief updating tasks. Starting with the simplest environment, we consider a setting where the state is binary and both states are equally likely. A decision-maker receives two simultaneous binary signals and reports her posterior about the state. Our analysis, conducted across various parameters, empirically documents how deviations from Bayesian predictions, termed mistakes, depend on signals' precisions. We observe two empirical regularities: first, participants make larger mistakes as the accuracy of both signals increases, even when the gap between signal accuracies remains fixed; second, participants make larger mistakes as the gap between signal accuracies widens. We show these regularities closely track non-linearities arising in Bayesian updating. The difficulty of integrating information from signals with different accuracies increases when Bayesian posteriors change significantly with small changes in signal accuracies. Regions in which these non-linearities are more pronounced are precisely the regions in which people make larger mistakes, unable to fully incorporate the extent to which one signal is more informative than the other.[2]

We explore several approaches that predict the two empirical regularities described above. The first approach directly links the difficulty of the updating task with the non-linearity of the Bayesian posterior. This approach has the advantage of measuring the difficulty based directly on task primitives without relying on a specific behavioral model.

[1] Not only humans employ Bayesian updating! Valone (2006) reviews experiments involving animals and concludes that a variety of them, across different ecological contexts, behave in a manner consistent with Bayesian updating.

[2] Throughout the paper, we associate mistakes participants make—the discrepancy between observed and Bayesian predicted posteriors—with task difficulty rather than task complexity. However, these two terms can be used interchangeably. The task becomes difficult due to the increase in complexity resulting from pronounced non-linearities.

The other two models we consider are behavioral. The first revisits Grether (1980), widely used in empirical studies of belief updating, and incorporates minor changes to include the aforementioned concepts. The second proposes an alternative model where an agent's inability to fully follow changes in the second derivative leads to only partial reactions to nonlinearities.

We next explore the relationship between task difficulty related to nonlinearities and alternative behavioral models predicting mistakes in updating, recently explored in the literature. The first suggests that the proximity of Bayesian prediction to corner belief (either 0% or 100%) is what drives mistakes. As we show in Section 4.2, participants have no issue arriving at the posteriors very close to the corner as long as they have only one signal to incorporate.[3] It is the presence of multiple sources of information that creates conditions for the task to become difficult. The second alternative explanation is the compression effect and its sensitivity to cognitive uncertainty (Enke and Graeber, 2023). We utilize the dataset collected by the above authors and show that controlling for cognitive uncertainty, the level of and gap between signal precisions remains a significant driver of participants' mistakes. We also demonstrate how our findings are different from a model of cognitive noise (Augenblick et al., 2023) as well as a model focusing on the cardinality of the state space (Ba et al., 2023). We find these results reassuring as they indicate that the difficulty of belief updating tasks related to nonlinearities is fundamentally distinct from other drivers of mistakes.

We next investigate how mistakes identified in scenarios with simultaneous signals translate to more typical belief-updating tasks, where the decision-maker receives information sequentially. To bridge this gap, we manipulate the information structure and information sequencing of the task. Information structure pertains to the idea that the same content of information can be presented in different ways. A scenario in which a decision-maker has an informative prior can alternatively be re-formulated as one in which they start with an uninformative prior but receive the equivalent amount of information through a signal. Although mathematically equivalent, these expositions differ conceptually. Our experiment provides the first empirical evidence of this equivalence. Regarding information sequencing, we alter the delivery of signals—either simultaneous or sequential—and, in the sequential scenario, whether the more accurate or less accurate signal is received first.

Our findings offer multiple insights. First, compared to what we term the Baseline

[3] In particular, we show that when participants start with a 50/50 prior and receive one partially informative signal, the vast majority arrive at the exact Bayesian posterior, regardless of how close this posterior is to the corner.

treatment—where participants begin with an informative prior, receive a signal, and afterward update their beliefs—providing the same information through two simultaneous signals significantly reduces errors. Notably, one of the parameterizations we examine uses parameters from the seminal paper by Kahneman and Tversky (1973), which has been employed in many subsequent papers. This parameterization is a leading example of base-rate neglect, one of the most well-documented and persistent biases observed in various settings (Benjamin et al., 2016; Benjamin, 2019; Esponda et al., 2023; Gneezy et al., 2013). According to our measure of difficulty, this parameterization is not considered difficult due to the relatively low precision of signals and the small gap between them. Our data reveals that presenting information through two simultaneous signals effectively addresses base-rate neglect, as the resulting posteriors are not statistically different from Bayesian predictions. This is an important result, as decades of research in both economics and psychology have shown that this bias is minimally affected by increased incentives (Gneezy et al., 2013), abundant feedback (Esponda et al., 2023), and the use of contextual representations (Gigerenzer and Hoffrage, 1995).

In our analysis of sequential information treatments, we find no aggregate impact on participants' reported beliefs when modifying the information structure. However, we document a significant recency bias that occurs regardless of task difficulty. We show that altering the sequence in which information is presented—either simultaneously or sequentially, and in the latter case, by varying the order of high and low accuracy signals—can reduce participants' errors. However, the most effective approach for disseminating information varies based on task difficulty. When task difficulty is low, simultaneous information release is most effective. When task difficulty is high, we leverage the recency bias to counteract the bias induced by task difficulty, making a sequential release of information more advantageous. Concluding our analysis, we use most of our treatments to decompose base-rate neglect and discover that it primarily arises from the sequencing of information and task difficulty.

Taken together, our experiments provide evidence that nonlinearities are influential in belief updating tasks. We believe that people's limited capacity to fully internalize nonlinearities extends beyond the realm of belief updating and presents an intriguing avenue for future research.

The remainder of the paper is structured as follows. We survey the literature in Section 1.1. In Section 2, we lay out the conceptual framework. We dedicate Section 3 to the experimental design and procedures. In Section 4, we present the aggregate results of our experiment. Individual level analisis are presented in Section A in the Appendix. We conclude in Section 5.

1.1 Literature Review

Our paper relates to and builds upon several strands of literature. We discuss these strands below and highlight our contribution in comparison to the prior findings.

Complexity literature. There is a rapidly developing and intriguing literature on decision complexity. The debate surrounding the definition of complexity and the empirical methods for identifying it are still in their early stages and largely dependent on context. In the domain of rules, complexity is found to be influenced by the number of states and transitions required to implement such a rule (Oprea, 2020; Banovetz and Oprea, 2022; Camara, 2021). In the domain of lotteries, complexity has been linked to the number of distinct outcomes in the lottery support (Bernheim and Sprenger, 2020; Puri, 2022; Fudenberg and Puri, 2022), the cognitive difficulty of aggregating outcomes and objective probabilities into a single value (Oprea, 2022), and excess similarity between lotteries in the choice set (Enke and Shubatt, 2023). In the inter-temporal choice problems, complexity has been linked to the difficulty of evaluating streams of future payments (Enke et al., 2023).

Closer to our setting are three recent papers exploring suboptimal updating through the lens of behavioral theories and experimental data. Enke and Graeber (2023) show a link between mistakes in updating and cognitive uncertainty measure, which captures how confident people are in their decisions. The data reveals that cognitive uncertainty is related to a compression effect of reporting beliefs closer to a 50-50 point. In addition, the authors manipulate the computational complexity of a task and find that cognitive uncertainty at least partly reflects the subjective perception of how difficult the problem is. Augenblick et al. (2023) study mistakes in belief-updating tasks with different signal precisions and find that people tend to over-infer from weak signals and under-infer from strong signals. The authors utilize a theory of cognitive imprecision about signal informativeness to accommodate both patterns. Ba et al. (2023) suggest that the difficulty of incorporating new information depends on the size of the state space and develops a behavioral model that incorporates two known psychological frictions, noisy cognition and representativeness. This model predicts under-reaction to new information when the state space is simple (consists of two states only), while the opposite is true when the state space is more complex.

In Section 4.2, we discuss at length all three above approaches, relate them to our notion of task difficulty, and explore the implications of these theories in the context of our experiment. To preview these results, we show that cognitive uncertainty, cognitive imprecision, and the cardinality of the state space cannot account for the mistakes we

observe across our treatments. We, therefore, view our work as complementary to the existing literature and proceed to formalize the connection between task difficulty and the nonlinearities embedded in Bayesian updating.

Finally, we note that our task difficulty notion relates to the literature that documents challenges and sub-optimal decisions apparent in environments with nonlinear features. Among the more prevalent patterns are the exponential growth bias, which is the tendency to underestimate compound growth processes prevalent in financial decisions (Wagenaar and Sagaria, 1975; Stango and Zinman, 2009; Levy and Tasoff, 2016, 2017), the schmeduling heuristics, which suggest simplified ways to construct mental representations of nonlinear incentive schemes (Rees-Jones and Taubinsky, 2020), and the difficulty in discounting atemporal payments that feature a large number of steps given the induced parameters (Enke et al., 2023).

Sub-optimal belief updating. There is a vast body of research in psychology and economics documenting errors in probabilistic reasoning that result in sub-optimal beliefs, i.e., beliefs that do not align with Bayesian predictions. The most relevant to our study are papers that document the effect information sequencing has on belief updating, e.g., primacy and recency effects (Pitz and Reinhold, 1968; Edenborough, 1975; Grether, 1992). Benjamin (2019) provides the most recent and comprehensive survey focusing on biases related to random samples and belief updating in general.

Base-rate neglect holds a prominent place in this literature as it is one of the more persistent phenomena and, therefore, one of the most frequently studied. First introduced by Kahneman and Tversky (1972) and Bar-Hillel (1980) and followed by many empirical papers scrutinizing this bias and theoretical models contemplating its origins (Benjamin, 2019; Benjamin et al., 2019). Of special interest is a recent paper by Esponda et al. (2023), which shows that even after ample opportunities to learn, base-rate neglect persists. The authors explore the main forces that hinder learning from feedback and find that mistakes that stem from misrepresentation of primitives of the environment are likely to be persistent. In this paper, we use base-rate neglect as our case study. However, we note that the insights we present are general and applicable to any setting involving probabilistic reasoning in the presence of new information.

What mitigates biased beliefs? Given the prevalence of biased beliefs and their importance in determining decisions, great efforts have been made to understand how responsive these biases are to various features of the environment and how to mitigate them.[4]

[4]See also the 'nudge' literature, which explores how to steer people into making better choices (Thaler and Sunstein, 2008; Thaler and Benartzi, 2004; Madrian and Shea, 2001).

Gneezy et al. (2023) study the role of financial incentives and find that base-rate neglect is largely unresponsive to stakes. Several papers document that the extent of base-rate neglect depends on whether the task is presented in terms of frequencies as opposed to probabilities (Koehler, 1996; Barbey and Sloman, 2007), whether the task is framed in an intuitive and contextual manner as opposed to an abstract way (Cheng and Holyoak, 1985; Gigerenzer and Hoffrage, 1995; Gneezy et al., 2023; Ganguly et al., 2000), and whether the task is presented as forecasting the future outcomes as opposed to updating beliefs about the state Fan et al. (2022). Esponda et al. (2023) find that information about primitives of the environment might hinder learning. Contrasting a treatment in which no primitives were provided with standard treatment, they find that in the former, elicited beliefs are eventually closer to the Bayesian posterior.[5]

We share with this literature the goal of understanding drivers of base-rate neglect and ways to mitigate it. While several manipulations from the papers discussed above reduce the bias to some extent, none come close to eliminating it. We examine a different manipulation and demonstrate that, for the typical parameters used in the literature, it eliminates base-rate neglect from the onset of the treatment.

2 Conceptual Framework

2.1 Setup

Consider a standard belief-updating task. The state is binary $\omega \in \{F, S\}$, denoting, for example, whether a project is a Failure or a Success. A decision-maker does not know the state but holds prior belief $P(F) = p_0$. They observe the realization of a signal s, which can take either a negative ($s = n$) or positive ($s = p$) value. The signal has accuracy θ_s, which summarizes the probability that it correctly reveals the state, $P(s = n|F) = P(s = p|S) = \theta_s$.

Upon observing a signal, Bayes' rule dictates that the updated beliefs in the form of a posterior-odds ratio can be written as

$$\frac{P(F|s)}{P(S|s)} = \frac{P(s|F)}{P(s|S)} \frac{P(F)}{P(S)} = \frac{P(s|F)}{P(s|S)} \frac{p_0}{1-p_0}, \tag{1}$$

[5]See also a stream of recent papers that discuss the relationship between incorrect mental models and task complexity (Enke and Zimmermann, 2019; Enke, 2020; Graeber, 2023). Worth highlighting is also Esponda et al. (2023) who, in one of their treatments, provide subjects with empirical frequencies of the joint distribution of signals and outcomes. Their findings suggest this approach can significantly mitigate the issue of base-rate neglect. However, it is important to note that the effectiveness of this solution is contingent upon gathering data over multiple rounds.

where $\frac{P(s|F)}{P(s|S)}$ is the base factor of the signal, and $\frac{p_0}{1-p_0}$ is the prior probability ratio. The posterior odds ratio underscores that both the prior and the signal contain valuable information, which the decision-maker uses to update their beliefs.

2.2 Information Structure

For our analysis, it is useful to distinguish between an informative and uninformative prior, where the latter is one that has minimal impact on the posterior. For our binary case, an uninformative prior is one that assigns equal probability to both states $p_0 = 1/2$. With this prior, once an agent receives a signal, their posterior is fully determined by the value and accuracy of this signal.[6] In this spirit, we intend to convey that the prior is uninformative: it does not interfere with the information from the signal, it is optimal to fully follow the signal. Henceforth, we call a prior *uninformative* if it assigns a probability of $1/2$ to each state and call it *informative* otherwise.

Consider an agent with an informative prior $P(F) = p_0 > 1/2$. If this agent receives no other information, their posterior-odds ratio will be $\frac{P(F)}{P(S)} = \frac{p_0}{1-p_0}$. We can reinterpret this prior as a posterior originating from an initial uninformative prior $P(F) = \tilde{p}_0 = 1/2$, and a signal with accuracy $\theta_s = p_0$ whose realized value is negative.[7] In this case, the posterior-odds ratio will be

$$\frac{P(F|s=n)}{P(P|s=n)} = \frac{P(s=n|F)}{P(s=n|S)}\frac{P(F)}{P(P)} = \frac{\theta_s}{1-\theta_s}\frac{1/2}{1/2} = \frac{p_0}{1-p_0}.$$

Thus, having a prior $p_0 > 1/2$ is equivalent to having an uninformative prior $\tilde{p} = 1/2$ and receiving a signal with accuracy p_0, that turned out to be negative. The mathematical requirements to calculate posteriors remain unchanged. We refer to this equivalence as *information structure*, pertaining to the idea that the same content of information can be presented in different ways.

This equivalence holds more generally, for instance, in cases in which an agent has an informative prior and receives additional information from the outset. Consider an agent with an informative prior $P(F) = p_0 > 1/2$ who receives a signal with accuracy θ_s. Once more, we can express the updated beliefs in the form of a posterior-odds ratio as in equation (1). Now, consider an alternative case where an agent receives two signals

[6] Specifically, if the accuracy of the signal is θ_s and the signal realization is negative (positive), the agent's posterior beliefs, as calculated by Bayes' rule, are that the project is a Failure (Success) with probability θ_s and a Success (Failure) with complimentary probability $1-\theta_s$. Thus, the new information fully determines their distribution of beliefs.

[7] Having $p_0 > 1/2$ is without loss of generality. If $p_0 < 1/2$, let the realized signal be $s = p$ and signal accuracy be $\theta_s = 1 - p_0$.

with accuracy θ_1 and $\theta_2 = \theta_s$, and has an initial prior $\tilde{p}_0$. Conditional on $s_1 = n$, the posterior-odds ratio will be

$$\frac{P(F|s_2, s_1 = n)}{P(S|s_2, s_1 = n)} = \frac{P(s_2|F)}{P(s_2|S)} \frac{P(s_1 = n|F)}{P(s_1 = n|S)} \frac{P(F)}{P(S)} = \frac{P(s|F)}{P(s|S)} \frac{\theta_1}{1-\theta_1} \frac{\tilde{p}_0}{1-\tilde{p}_0}.$$

In the special case in which the accuracy of the first signal is $\theta_1 = p_0$, and the prior is uninformative, $\tilde{p}_0 = 1/2$, the above reduces to

$$\frac{P(F|s_2, s_1 = n)}{P(S|s_2, s_1 = n)} = \frac{P(s|F)}{P(s|S)} \frac{p_0}{1-p_0} \frac{1/2}{1/2}. \tag{2}$$

Note that equation (2) is equal to the posterior-odds ratio under the initial problem, equation (1).[8] Thus, the two information structures, one with an informative prior and a signal and another with an uninformative prior and two signals, are mathematically equivalent. Importantly, this equivalence holds regardless of whether the signals arrive simultaneously or sequentially. This is critical for our setup because, as we move from one treatment to another, we alter the structure and sequencing of information, yet the core mathematical problem remains unchanged.

2.3 Task Difficulty

Motivating Example. To illustrate the general idea behind our notion of task difficulty, consider the following example. There are two information structures, both with an uninformative prior and two signals, the accuracy of which differs by five percentage points. In the first case, signal accuracies are $\theta_1 = 0.85$ and $\theta_2 = 0.80$, while in the second, they are $\theta_1 = 0.97$ and $\theta_2 = 0.92$. Consider an individual who observes two signals, $s_1 = n$ and $s_2 = p$, and updates posterior beliefs. If both signals had identical accuracy, say both equal to 0.80, or both equal to 0.92, then it seems natural to place equal weight on both signals and form a posterior equal to the original prior of 50%. However, because the first signal is more accurate than the second, individuals may argue they should weigh the first signal slightly more and, thus, compute a posterior that leans more toward Failure than Success.

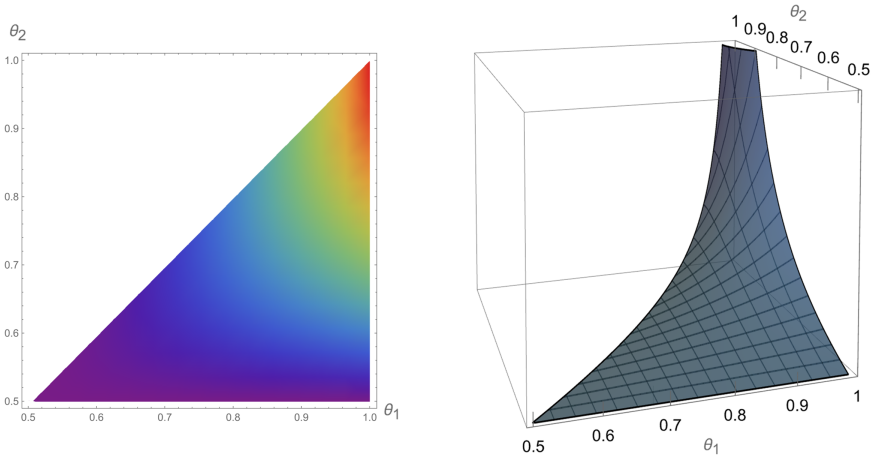
Although this is true in both cases, the extent to which the first signal is more informative than the second is quite different between them. The Bayesian posterior is 41.3% in the first case, seemingly a plausible value near which individuals' average posteriors may lie. In contrast, the Bayesian posterior is 26.2% in the second case, despite the dif-

[8] $\frac{P(s|F)}{P(s|S)}$ will be equal to $\frac{\theta_s}{1-\theta_s}$ or $\frac{1-\theta_s}{\theta_s}$ depending on the realized value of the signal.

ference in signal accuracies remaining only five percentage points. This pattern becomes even more pronounced with higher accuracy signals or as the disparity between signal accuracies increases.

Task Difficulty and Nonlinearities. The difficulty illustrated above stems from the non-triviality of incorporating information from signals with different accuracies. The challenge is to know how much more one needs to react to the higher accuracy signal compared to the lower accuracy one. The answer to this question depends on the accuracy of both signals. We argue individuals may struggle to fully internalize the extent to which Bayesian updating requires non-linear thinking. Figure 1 depicts the derivative of the posterior with respect to the high-accuracy signal.[9] The marginal increase is unchanged only when facing a single signal ($\theta_2 = 1/2$ reduces the low-accuracy signal to an uninformative one). In all other cases, the marginal change in the posterior depends not only on the signal's own accuracy but on the accuracy of the other signal as well, i.e., the Bayesian posterior is no longer linear. It is these nonlinearities in Bayesian updating that we believe contribute to the varying levels of difficulty.

Figure 1: Bayesian Posterior Derivative w.r.t high accuracy signal, θ_1



Notes: The left graph depicts a heatmap (warmer colors represent higher values), and the right graph is 3D plot. In both graphs, we focus on $\theta_1 \geq \theta_2$, since θ_1 denotes the high accuracy signal.

We argue that if individuals fail to fully account for these nonlinearities, they will perform better with belief updating in regions where the nonlinearities are less pronounced. Conversely, when these nonlinearities are high—a small change in signal precision leads to a large change in posteriors—we argue that individuals' mistakes will be larger. Referring back to the example above, the case with $\theta_1 = 0.85$ and $\theta_2 = 0.80$ falls in a region

[9] Predictions are unchanged if alternatively, we focused on the low accuracy signal.

where nonlinearities are less pronounced compared to when $\theta_1 = 0.97$ and $\theta_2 = 0.92$. Therefore, we predict that individuals would make larger mistakes in the latter case.

Concrete Formulations. While there are several ways to capture this idea formally, in the Appendix we offer a parsimonious approach as well as two alternative models that yield qualitatively similar predictions to those described here. First, in Section B.1, we offer an approach based only on Bayesian posteriors, in which the difficulty of the problem increases in the nonlinearity of the posterior. Next, in Section B.2, we explore a modification of the Grether (1980) model, which has been extensively used in empirical work studying belief updating, and make minor restrictions incorporating the aforementioned concepts. This modified model is more inert than the Bayesian model, leading to large differences, especially in regions where large changes in posteriors are expected for small changes in signal accuracies. Finally, in Section B.3, we explicitly model an agent's inability to fully follow changes in the second derivative and, thus, only partially react to nonlinearities.

The frameworks mentioned above represent some of possibly many approaches aiming at highlighting the idea that nonlinearities involved in Bayesian updating may influence an individual's ability to carry out proper belief updating. We suspect—and hope—that future work will identify and test even more specifications that speak to this phenomenon.

Testable Implications. The above discussion, regardless of the specific formalization we consider, leads to the following testable implications.

1. Task difficulty increases as the level of signal accuracies increases, holding fixed the gap between signal accuracies.

2. Task difficulty increases as the gap between signal accuracies increases, holding fixed the level of signal accuracies.

It is these testable implications that we take to the data. As we describe in detail in the next section, we utilize six parameterizations, across which we systematically vary both the absolute level of signal accuracies and the gaps between them. The empirical footprint of these testable implications is the wedge between posteriors reported by participants and Bayesian predictions. If our notion of task difficulty holds water, we expect to see greater mistakes in more difficult belief-updating problems—higher levels of and gaps between signal accuracies.

3 Experiment Design

We designed our experiment to manipulate three key aspects: task difficulty, information structure, and information sequencing.

To capture the impact of task difficulty, we vary parameter values. To capture the impact of information structure, we vary whether participants receive information through an informative prior and one signal or through an uninformative prior and two signals. Finally, to capture the impact of the sequencing of information delivery, we vary whether the two signals arrive simultaneously or sequentially. To decompose any possible interaction between the timing of a signal's arrival and its accuracy, we have two sequential information arrival treatments in which we vary the order of signal accuracies: high accuracy followed by low accuracy and vice versa. Below, we first describe the main task participants encountered in each treatment and then provide details about the structure and parameters of each treatment.

Main Task. In each treatment, participants face a standard belief-updating task with a binary state and binary signal(s). At the outset of each round, the state is represented by a project selected from a pool of projects; each project has a p chance of being a Failure and a $1-p$ chance of being Successful. Participants know p but do not know whether the project is successful or not. Depending on the treatment, participants receive either one or two conditionally independent signals with known accuracies.[10] In treatments with two signals, participants receive signals with different accuracies. Let θ_1 (θ_2) denote the accuracy of the higher (lower) accuracy signal. We use the strategy method and elicit participants' posterior beliefs for each possible signal realization.[11] Relying on the strategy method ensures we collect a balanced dataset. Each participant participates in only one treatment and plays a total of 20 rounds.[12]

Feedback. At the end of each round, participants are informed about the realized value of the signal(s) (positive or negative) and state (Success or Failure). Realized signal and state values from all previous rounds are stored at the bottom of the screen in an easy-to-read table. We include detailed feedback to mitigate potential memory issues that may disrupt learning and confound results; see screenshots in the Online Appendix.

[10] Signal accuracy is the probability that the signal correctly reveals the state: $P(s = p|S) = P(s = n|F) = \theta_s$.

[11] The strategy method is a common practice in many experiments, including beliefs-updating experiments (Gneezy et al., 2013; Esponda et al., 2023). For the comparison between the strategy method and the direct response method, see Brandts and Charness (2011).

[12] Repetition of a task is a standard technique in experiments, which allows participants to adjust to the interface and further arrive at their optimal response. We address learning in the data analysis section.

Treatments. In the **Baseline** treatment, participants have an informative prior p and receive one signal with accuracy θ_2.[13] This treatment mimics the classic experiment of Kahneman and Tversky (1972) and the follow-up literature on base-rate neglect (Benjamin, 2019; Esponda et al., 2023).

In the **Simultaneous** treatment, participants have an uninformative prior and receive two signals simultaneously. The signals are conditionally independent and have accuracies θ_1 and θ_2. With two binary signals, there are four possible combinations of signal realizations. However, since the prior is uninformative, the case in which both signals are positive is the mirror equivalent of that in which both signals are negative. Similarly, the first signal being negative and the second being positive is the mirror of the case in which the first signal is positive and the second is negative. In other words, asking four questions would have been redundant. To keep treatments comparable, we randomly draw a value for the first signal and rely on the strategy method to allow for both a positive and a negative realization of the second signal. For more details, see the Online Appendix.[14]

In the **Sequential High-Low** treatment, participants have an uninformative prior and receive two signals. Unlike the Simultaneous treatment, these signals arrive sequentially. Participants receive the higher accuracy signal first. After observing the first signal realization, participants submit their updated beliefs about the state. Afterward, relying on the strategy method, participants submit their beliefs for two possible realizations of the second signal.[15] Thus, beliefs are elicited twice, once after the first signal realization and once via the strategy method for both possible values of the second signal.

Finally, the **Sequential Low-High** treatment is identical to the Sequential High-Low treatment except that participants receive the lower accuracy signal first.

In summary, for each treatment in each round, we elicit two beliefs: one when the signal aligns with the prior (or the first and second signals align) and another when the signal is misaligned with the prior (or the first and second signals are misaligned).[16]

Parameters. Table 1 summarizes our treatments and parameters. We use two main sets of parameters, denoted below by A and B. Under parametrization A, in treatments in

[13] We vary whether p represents the probability of Success or Failure. This assignment is determined randomly, with equal probability, for each participant at the beginning of the experiment and remains unchanged throughout.

[14] In addition to guaranteeing a balanced dataset, this design ensures that participants encounter all possible combinations of positive and negative signal realizations throughout the rounds.

[15] This makes the design comparable with the Simultaneous treatment in which the first signal value is drawn, whereas beliefs are elicited for both possible values of the second signal.

[16] We say the signal is aligned with the prior if the prior leans towards Failure (Success) and the realized signal value is negative (positive). In treatments with two signals, signals are aligned (misaligned) if both have the same realized value (one is positive and the other is negative).

which participants receive two signals, signal accuracies are set to $(\theta_2, \theta_1) = (0.80, 0.85)$, whereas in the Baseline treatment (denoted by $\tilde{A}$), the prior accuracy is set to $p = 0.85$ and signal accuracy to $\theta_2 = 0.80$. This is the classic set of parameters used in the base-rate neglect literature (Kahneman and Tversky, 1972; Benjamin, 2019; Esponda et al., 2023). Under parametrization B, in treatments in which participants receive two signals, signal accuracies are set to $(\theta_2, \theta_1) = (0.85, 0.95)$, whereas in the Baseline treatment (denoted by $\tilde{B}$), the prior accuracy is set to $p = 0.95$ and signal accuracy to $\theta_2 = 0.85$. As described in Section 2.2, within each parameterization, the Bayesian predictions are identical for all treatments. The primary purpose of parametrization B is to assess the validity of the task difficulty notion described in Section 2.3, and to act as a test of robustness for the findings from the initial parametrization A.

Table 1: Sessions, Treatments, and Parameter Values

Session	# Participants	Treatment	Parameter	Prior	Low Accuracy Signal	High Accuracy Signal
1	101	Baseline	$\tilde{A}$	p=0.85	$\theta_2 = 0.80$	—
2	101	Simultaneous	A	p=0.50		$\theta_1 = 0.85$
3	101	Sequential High-Low				
4	100	Sequential Low-High				
5	99	Baseline	$\tilde{B}$	p=0.95	$\theta_2 = 0.85$	—
6	99	Simultaneous	B	p=0.50		$\theta_1 = 0.95$
7	102	Sequential High-Low				
8	100	Sequential Low-High				
9	99	Simultaneous	C	p=0.50	$\theta_2 = 0.75$	$\theta_1 = 0.85$
10	100		D		$\theta_2 = 0.80$	$\theta_1 = 0.90$
11	100		E		$\theta_2 = 0.85$	$\theta_1 = 0.90$
12	100		F		$\theta_2 = 0.90$	$\theta_1 = 0.95$

Four additional parameterizations, C, D, E, and F, allow us to decompose how the level and gap between signal accuracies are linked to task difficulty. Based on our conceptual framework presented in Section 2.3, we can rank all six sets of parameters in terms of difficulty, with parameterization A being the easiest and parameterization B being the most difficult. The remaining sets can be ranked based on the two organizing principles, where C_j denotes the difficulty of parameterization j:

1. Task difficulty, as described in Section 2.3, increases as the level of signal accuracies increases, holding constant the gap between them:
   - $C_A(0.80, 0.85) < C_E(0.85, 0.90) < C_F(0.90, 0.95)$
   - $C_C(0.75, 0.85) < C_D(0.80, 0.90) < C_B(0.85, 0.95)$

2. Task difficulty increases as the gap between signal accuracies increases, holding constant the level:

   - $C_A(0.80, 0.85) < C_D(0.80, 0.90)$
   - $C_E(0.85, 0.90) < C_B(0.85, 0.95)$

Subject Pool. We conducted our experiment on the Prolific platform with roughly 100 participants in each of the 12 treatments, for a total of 1202 participants. We recruited participants between the ages of 18 and 70, who were living in the United States, were fluent in English, and had a high approval rating on Prolific. For each treatment, an equal number of men and women were recruited. The main experiment was carried out in October - December 2022, while two additional treatments (D and E) were conducted in July 2023.

Participants' Payments. In all treatments, participants received a $5 payment upon completion. In addition, each participant had a 20% chance to be selected into a bonus group. For the selected participants, one of the experimental rounds was randomly chosen for payment. The answers submitted in the chosen round determined whether the selected participant received an additional bonus of $20. We used the standard BDM method to incentivize subjects to truthfully state their beliefs.[17] The experiment lasted 20 minutes on average, and participants earned, on average $7.97.

Implementation. The experiment was approved by Caltech (IR22-1237) and Duke University IRB (2023-0033) and preregistered on aspredicted.org.[18] The experimental software was programmed in oTree Chen et al. (2016). Instructions and screenshots of the interface are presented in the Online Appendix.

[17] The BDM is theoretically an incentive-compatible method for eliciting truthful responses regardless of participants' risk attitudes Becker et al. (1964). In addition, to help participants understand this method, we told them that they had no incentive to report beliefs falsely if they wanted to maximize the expected payoff in the experiment. Danz et al. (2021) shows that announcing that truth-telling is optimal is an effective way to elicit true beliefs.

[18] The experiment was conducted in three waves: the initial wave with parametrization A and B in October 2022; decomposing task difficulty treatments, parametrization C and F, in December 2022; additional task difficulty treatments, parametrization D and E, in June 2023. Each wave was separately preregistered on aspredicted.org; see preregistration 1, 2, and 3.

4 Results

We begin our analysis with the Simultaneous treatment—an environment free of the effects of structure and sequencing. It is in this environment that we study our notion of task difficulty. The main analysis of task difficulty is presented in Section 4.1, where we utilize all six parameterizations of the Simultaneous treatment. In Section 4.2, we go through a variety of robustness checks supplemented by data from Enke and Graeber (2023) for an additional assessment.

Having established our findings with regard to our notion of task difficulty, we turn to the Baseline and Sequential treatments. In Section 4.3 and Section 4.4, we focus on parameterization A ($\tilde{A}$) and B ($\tilde{B}$) where we analyze the effect of the structure and sequencing of information delivery. In Section 4.5, we use the discrepancy arising from signal sequencing to counter the discrepancy arising from task difficulty. Finally, in Section 4.6, we quantify the extent to which information structure, sequencing, and task difficulty affect belief updating across the least and most difficult parameterizations A ($\tilde{A}$) and B ($\tilde{B}$), respectively. The individual-level analysis is presented in Section A in the Appendix. For the remainder of the paper, with a slight abuse of notation, we will refer to parameters $A(\tilde{A})$ and $B(\tilde{B})$ as simply parametrization A and B, respectively.

Approach to data analysis. We focus our analysis on cases in which participants receive misaligned information.[19] We do so because when participants receive aligned information, the Bayesian posterior probabilities are very close to zero.[20] With predicted values so close to the 0 border, implementation errors participants may have are unlikely to be mean zero. We regard utilizing this data for our main analysis as less than ideal. Thus, as stated in our preregistration, our focus is on elicitations from misaligned signals.[21] Nonetheless, we utilize all elicitations when conducting individual-level analysis in Section A and illustrate them in the Appendix.

To simplify the presentation and eliminate redundancies, in our data analysis, we normalize the prior and the high-accuracy signal to be negative. Since we focus on the elicited beliefs from misaligned signals, this normalization implies a positive value for the low-accuracy signal.

[19] Recall, in the Baseline treatment, information is aligned (misaligned) if the signal's realization agrees (disagrees) with the direction in which the prior leans. In all other treatments, information is aligned (misaligned) if the realized values of the signals are the same (different).

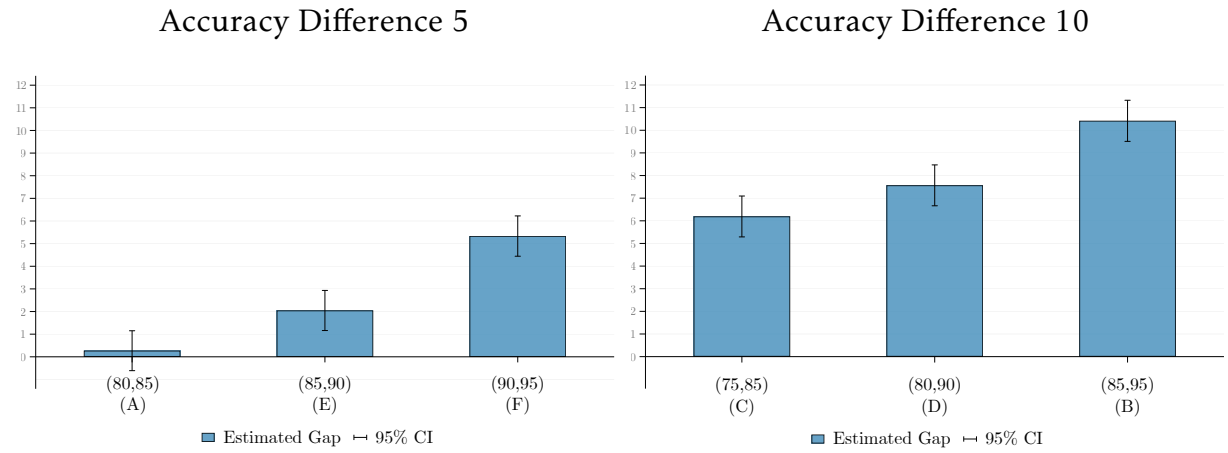
[20] These probabilities are 0.042 for parametrization A, 0.009 for B, 0.055 for C, 0.027 for D, 0.019 for E, and 0.006 for parametrization F.

[21] An alternative approach would be to treat these elicitations as truncated. However, doing so requires making assumptions about the nature of the truncation and the distribution of implementation errors. These assumptions would naturally not be without loss of generality.

4.1 Task Difficulty

In this section, we present experimental results from the Simultaneous treatments. We present the disparity between the observed and Bayeisan posteriors, which we call mistakes for brevity. Figure 2 displays how mistakes respond to an increase in the level of signal accuracies, keeping the difference between the two signals' accuracies fixed. In the left panel, the difference between signal accuracies is 5, whereas in the right panel, the difference is 10. Our data confirms that regardless of the difference between signal accuracies, an increase in the level leads to larger mistakes, i.e., the gap between reported posteriors and Bayesian predictions.

Figure 2: The Impact of Signal Accuracy Levels on Task Difficulty

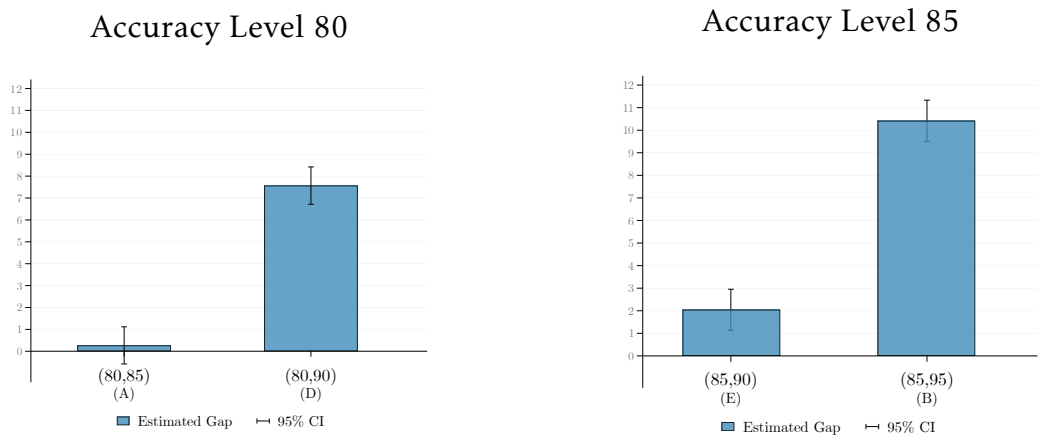


Notes: We report the difference between the observed and Bayesian posteriors, averaged across participants in all 20 rounds, alongside the 95% confidence intervals, clustered at the individual level. The horizontal axis depicts the accuracy of the low-accuracy signal and lists the label of each parameterization below.

Figure 3 shows how mistakes respond to an increase in the difference between signal accuracies, keeping the level fixed. In the left panel, the signal accuracy level is fixed at 80, whereas in the right panel, the level is 85. Our data confirms that regardless of the level of signal accuracies, an increase in the difference results in an increase in mistakes, i.e., the gap between observed and Bayesian posteriors.

We collect these findings in Table 2, where we present a regression of the gap between the reported and the Bayesian posterior on a constant, the level, and the difference between signal accuracies. The first column of Table 2 utilizes the whole dataset from the six treatments, whereas the second column relies on data from the last five rounds. The regression confirms the observation from the graphs, where we see that both the difference and the level of signal accuracies have a sizable and statistically significant effect.

Figure 3: The Impact of Signal Accuracy Difference on Task Difficulty



Notes: We report the difference between the observed and Bayesian posteriors, averaged across participants in all 20 rounds, alongside the 95% confidence intervals, clustered at the individual level. The horizontal axis depicts the difference in signal accuracies and lists the label of each parameterization below.

Table 2: Accuracy level and Difference impact on observed Gap

	Gap	
	All Rounds	Last 5 Rounds
Difference	1.566***	1.555***
	(0.240)	(0.289)
Level	0.464***	0.421***
	(0.124)	(0.152)
N (observations)	11980	2995
K (individuals)	599	599

Individual-level clustered errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Result 1 (Mistakes: Level and Difference)**.** *The level of and difference between signal accuracies increase the gap between observed and Bayesian posteriors.*

Linear Thinking Estimation. In this section, we estimate one of the models discussed above, according to which an agent can only partially incorporate nonlinearities. As we show in detail in Section B.3 in the Appendix, the posterior of such an agent can be decomposed into two parts: the Bayesian posterior and a fully linear posterior:

$$\tilde{\pi}(S|s_2 = p, s_1 = n) = \alpha \underbrace{\frac{\theta_2(1-\theta_1)}{\theta_2 + \theta_1 - 2\theta_2\theta_1}}_{\text{Bayesian Posterior}} + (1-\alpha)\underbrace{\left(\frac{1}{2} - \theta_1 + \theta_2\right)}_{\text{Fully Linear}}.$$

The α parameter quantifies the agent's ability to incorporate nonlinearities, where a value of $\alpha = 0$ indicates a complete failure to incorporate nonlinearities in belief updating, while $\alpha = 1$ signifies behavior indistinguishable from Bayesian updating. We estimate α

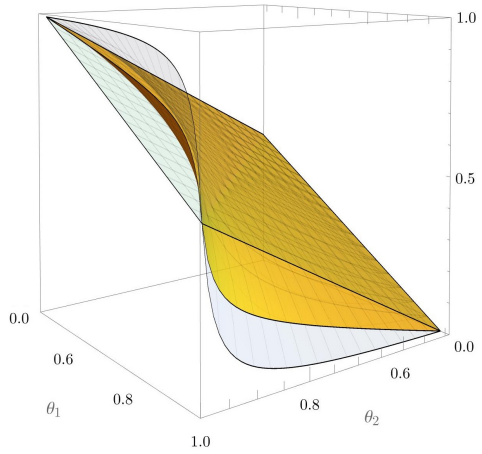
via linear regression and present the results in Table 3. In the first column, we analyze data pooled from all 20 rounds, while the second column focuses on data from the last five rounds. The graph of the estimated model is presented in Figure 4.

Table 3: Estimated α

	All Rounds	Last 5 Rounds
$\hat{\alpha}$	0.417***	0.564***
	(0.0555)	(0.0663)
N (observations)	11980	2995
K (individuals)	599	599

Individual-level clustered errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 4: Estimated Model



Notes: The figure illustrates the Bayesian ($\alpha = 1$) and fully linear ($\alpha = 0$) models through transparent graphs, along with the estimated model ($\alpha = 0.56$) via the yellow graph.

The estimated value of α is approximately 0.42 (0.56) when considering the entire dataset (the last five rounds).[22] Therefore, the model that best fits the data suggests that participants lie between Bayesian and linear updaters, indicating that participants face challenges in properly incorporating nonlinearities while still exhibiting some degree of nonlinear thinking.

Result 2 (Updating: Bayesian vs Linear)**.** *Participants are capable of incorporating nonlinearities only partially. Behavior is best described by a model that roughly lies between Bayesian and linear updating.*

4.2 Task Difficulty Robustness Checks

Task Difficulty and Cognitive Uncertainty We believe it is useful to compare the implications of our task difficulty measure with those predicted by the cognitive uncertainty notion studied in a recent paper by Enke and Graeber (2023). Cognitive uncertainty has been linked to a compression effect, according to which individuals tend to report beliefs closer to the middle value of 50 when they are more uncertain of their answers. Enke and

[22] By conducting a similar exercise and employing a nonlinear regression on the alternative modified Grether model, analyzed in Section B.3, we obtain an estimated value of approximately 0.46 (0.58) when utilizing the entire dataset (the last five rounds).

Graeber (2023) provides empirical support for this idea by gathering a new dataset consisting of belief elicitations as well as measures of cognitive uncertainty at the individual level.

We utilize data from Enke and Graeber (2023).[23] Aiming for comparability with our dataset, we examine cases where participants have an informative prior and receive a single signal.[24] Regression analysis, presented in Table 4, demonstrates that the level of and difference between the accuracies of information sources, which represent our measure of task difficulty, significantly impact the gap between reported and Bayesian posteriors even after controlling for cognitive uncertainty.

Table 4: Robustness Check

	Gap					
				Low CU	Mid CU	High CU
Difference	0.168***	0.173***	0.178***	0.118***	0.247***	0.224***
	(0.0154)	(0.0153)	(0.0156)	(0.0220)	(0.0252)	(0.0402)
Level	0.234***	0.232***	0.226***	0.201***	0.246***	0.329***
	(0.0495)	(0.0494)	(0.0491)	(0.0746)	(0.0755)	(0.112)
Cognitive Unc		0.0569***	0.0529***	0.0693	0.144**	0.0227
		(0.0173)	(0.0176)	(0.0587)	(0.0620)	(0.120)
Other Controls	No	No	Yes	No	No	No
N	2866	2866	2866	1496	1003	367

Individual-level clustered errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Each column represents a separate regression of the gap between reported and Bayesian posteriors. *Other Controls* include participants' age, education, Raven scores, and gender. We renormalize $Difference = \frac{Difference}{\max(Difference)} \cdot 100$ and $Level = \frac{Level}{\max(Level)} \cdot 100$ to ensure they are in the same order of magnitude as Cognitive Uncertainty, which can take values between 0 and 100. The *Low*, *Mid*, and *High CU* columns group observations with $CU \leq 33$, $33 < CU \leq 66$ and $66 < CU$ respectively.

In the last three columns, we segment the data by low, medium, and high cognitive uncertainty scores and present separate regressions for each. In these cases, cognitive uncertainty loses power due to restriction to specific subsets. Our takeaway from these additional regressions is that, regardless of the controls and the subsets defined by participants' cognitive uncertainty, our main parameters of interest remain both statistically significant and substantial in magnitude.

[23] We are grateful to the authors for providing us with their data.

[24] The parameters utilized have prior values of $(50, 70, 90, 95, 99)$ and signal accuracies of $(65, 70, 75, 90)$. Our findings in Section 4.4 indicate that beliefs observed in the case with an informative prior and one signal should closely resemble those obtained from a treatment involving sequential information arrival with two signals and an uninformative prior. The sequential arrival of information is less than ideal for the study of task difficulty, as sequencing also impacts elicited beliefs. However, in Section C.2 in the Appendix, utilizing symmetric parameter cases, we perform a back-of-the-envelope calculation and find that, given these parameter values, sequencing can account for only a small part of the increase in the observed gap, with the majority being attributed to the increase in task difficulty.

Our interpretation of these results is that our notion of task difficulty and cognitive uncertainty are distinct phenomena that can coexist and jointly influence the posteriors in belief-updating tasks. In other words, our measure is not necessarily linked to perceived uncertainty: a participant may be very confident and very wrong at the same time.[25]

Task Difficulty and Proximity to Corner Beliefs Recall that in the sequential treatments, participants' beliefs are elicited twice, once after receiving the first signal and once more via the strategy method conditional on the second signal. Further, recall from our discussion of task difficulty in Section 2.3 that when faced with an uninformative prior and a single signal, the Bayesian posterior is linear. Such tasks are, according to our measure, not difficult. We next look into this prediction.

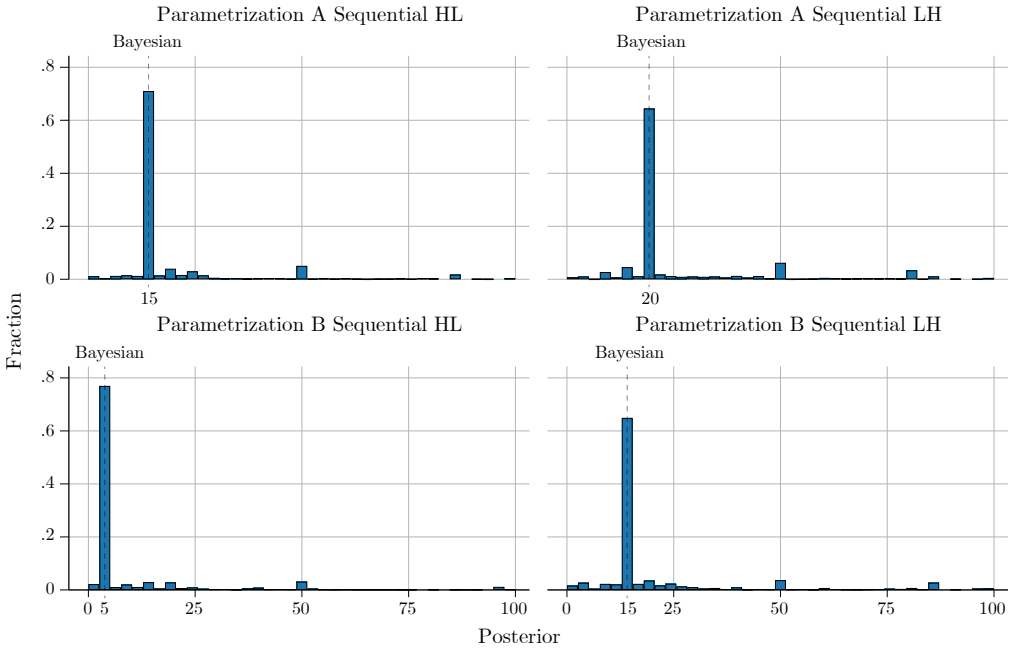
Utilizing sequential treatments, our data shows that barring implementation noise, participants' reported beliefs after receiving the first signal are largely in line with Bayesian posteriors, see Figure 5. The upper (lower) graphs represent parametrization A (B), with dashed lines indicating the Bayesian posterior centered at 15, 20, 5, and 15, respectively. The vast majority of participants' choices correspond to these levels. Because noise to the left is bounded by 0, whereas the boundary for noise to the right is much further, the estimated averages tend to be a few percentage points higher than the accuracy of the signal, whereas the median values of the posteriors match the Bayesian posteriors in all four cases. Therefore, when confronted with an uninformative prior and a single informative signal, participants accurately estimate posterior beliefs.[26] This aligns with our notion of task difficulty and the linear nature of this particular case.

Importantly, Figure 5 reveals that participants are capable of making correct choices far from the middle (50%) point, e.g, in the most extreme case, when the correct choice is 5 in the Sequential B HL treatment, almost 80% of participants make this choice. Thus, our documented increase in mistakes as signal precision increases can not be attributed to participants' inability to make correct choices near the extremes, as the graph shows they are more than capable of doing so.

[25] Interestingly, running a regression of cognitive uncertainty on the difference and level of signal precisions leads to an insignificant estimated value on the level and a negative and statistically significant estimated value on the difference. That is, participants reported uncertainty does not appear to rise with increased signal precision. On the contrary, reported uncertainty tends to *decrease* as the difference between signal precisions increases.

[26] Two deviations worth mentioning are a small share of individuals choosing a posterior of 50 and a small share of individuals making the inverse of the correct choice: 85 instead of 15, 80 instead of 20, 95 instead of 5, and 85 instead of 15, respectively. A deeper dive into the data reveals that these are occasional mistakes a few participants make rather than systematic deviations.

Figure 5: Sequential Treatments - Posteriors after the First Signal



Notes: Above, we report the histogram of participants' posteriors. The vertical axis represents the fraction of choices, whereas the horizontal axis corresponds to the particular posterior.

Task Difficulty and Cognitive Noise Augenblick et al. (2023) demonstrate that individuals may under- or over-infer from signals in a manner consistent with a cognitive noise model, where noise interacts with signal precision. We see these findings as aligned with the message of our paper, though there are significant differences. In our framework, not only does the precision of the particular signal matter, but the number of relevant information sources and the precision of other signals are of primary importance. For example, if we estimate the model from Augenblick et al. (2023)

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) = \log\left(\frac{\pi_0}{1-\pi_0}\right) + k\log\left(\frac{p}{1-p}\right)^{\beta}$$

in the baseline treatment, under parameters A and B, our estimates are $\hat{k} = 1.35$ and $\hat{\beta} = 1.78$. Recall that for a Bayesian, both values are 1. As we will soon see in Section 4.3, in such cases, participants excessively overfollow the signal. Consistent with these observations, the estimation indicates that participants excessively overfollow the signal, both on average and even more when the signal has higher precision. We next estimate the same parameters under the sequential treatment. To ensure an environment with a prior and one signal, we focus on the updated beliefs after receiving the first signal. Estimating the model again yields $\hat{k} = 0.79$ and $\hat{\beta} = 1.09$, which are much closer to the Bayesian values. The additional weight from $\hat{\beta} > 1$ is mainly offset by the lower weight from $\hat{k} < 1$, making participants' choices remarkably close to the optimal posteriors, as discussed in **Task Difficulty and Proximity to the Corner Beliefs** above.

These large differences in our estimated parameters indicate that if we were to use the estimated model from one setting to predict behavior in another, we would do a rather poor job. Importantly, this implies that the behavior displayed by participants depends on the entire learning environment and not only on the precision of the signal in isolation. In this paper, we argue that the number of relevant information sources and the precision of *all* these sources determine whether individuals effectively incorporate information.

Task Difficulty and the Cardinality of the State Space Of relevance is also the work of Ba et al. (2023), who demonstrate that how individuals incorporate signals depends on the complexity of the state space. They find strong evidence that the number of realizations a signal can have greatly influences its incorporation. Our work differs by showing that even when the number of information sources and the cardinality of possible outcomes are fixed, individuals' performance in belief updating tasks varies greatly. Specifically, we demonstrate that participants perform poorly in highly nonlinear regions but are indistinguishable from Bayesians in highly linear regions. Thus, while the cardi-

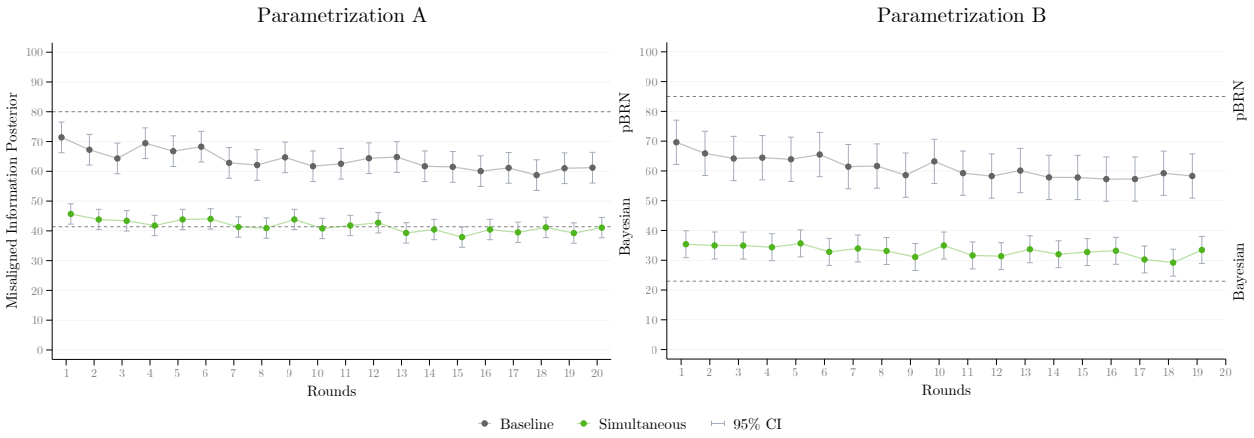
nality of the state space is of crucial importance, as demonstrated by Ba et al. (2023), we emphasize that significant behavioral variation is possible even with a fixed cardinality.

To summarize, we view the highlighted papers as complementary to our work, as we all aim to understand what features affect belief updating. While we share commonalities with previous research, we demonstrate that what we capture is different from cognitive uncertainty and cognitive noise, is not driven by the proximity of corner beliefs, and affects belief updating even if the cardinality of the state space is kept fixed. We emphasize that task difficulty depends not only on signal precision but also on the number of relevant information sources and the precision of all sources.

4.3 Baseline vs Simultaneous Treatments

In Figure 6, we present the round-by-round average posteriors under misaligned information. The two dashed lines represent the Bayesian posterior (41.38 in *A* and 22.97 in *B*) and the posterior for an agent exhibiting perfect base-rate neglect (80 in *A* and 85 in *B*). The latter refers to an agent that completely disregards the information from the prior and relies solely on the signal to form a posterior.

Figure 6: Posteriors in Baseline and Simultaneous A and B Treatments



Notes: We report the round-by-round average posteriors under misaligned information, alongside the 95% confidence intervals, clustered at the individual level. The lower horizontal dashed line depicts the Bayesian posterior, while the top dashed line depicts the perfect base-rate neglect posterior.

We start with parameterization *A*, which represents the low-difficulty setting according to our notion presented in Section 2.3. This parameterization is also the leading parameterization used in both Economics and Psychology literature to study how people update their beliefs (Benjamin, 2019), and, specifically, documenting the base-rate ne-

glect phenomena. In the Simultaneous treatment, both signals are received at the same time, eliminating any effects of information sequencing. Moreover, the effect of information structure is minimized as all available information is conveyed through signals alone, without the presence of an informative prior. Therefore, the Simultaneous treatment under parametrization A is the natural initial comparison to the Baseline treatment as it minimizes the effect of information structure, information sequencing, and task difficulty.

Throughout all rounds, we see minimal learning in the Baseline treatment, maintaining an average posterior of 63.69. These features are consistent with the existing literature.[27] This characteristic of minimal learning holds for all treatments in our experiment. In the Simultaneous treatment, however, the average belief is 41.65 and not statistically distinguishable from the optimal Bayesian level (see Table 5); the p-value of the difference is 0.786. We find this observation striking, considering that, as emphasized in Section 2.2, the mathematical problem underlying both treatments is identical. Moreover, it is worth noting that this complete correction occurs right from the outset.

We consider this to be an important finding, especially given the comprehensive body of work in both Psychology and Economics that establishes base-rate neglect as one of the most persistent deviations from Bayesian updating, showing minimal responsiveness to higher incentives Gneezy et al. (2023), ample feedback Esponda et al. (2023), or the use of more relatable contexts over mathematical constructs Gigerenzer and Hoffrage (1995). Certain approaches that have managed to mitigate deviations from optimal updating involve the utilization of aggregate statistics computed over several rounds. However, in real-life situations, individuals may not encounter the same problem repeatedly, and aggregating private information can be challenging. Thus, we consider it noteworthy that appropriate structuring and timing of information alone can effectively aid in, or in this case, completely rectify belief updating.

Result 3 (Belief Updating Immediate Correction)**.** *In low-difficulty environments, delivering information through two simultaneous signals leads to an estimated mean statistically indistinguishable from the Bayesian posterior.*

Turning to the right panel of Figure 6 which presents the high-difficulty setting, parameterization *B*, we note that the Simultaneous treatment no longer achieves the optimal Bayesian level. The average of elicited posteriors across the 20 rounds is 33.39, whereas the Bayesian posterior is 22.97. With both information structure and sequencing being

[27] See, for example, Esponda et al. (2023) who have the same baseline as us with the same parameter values but run the experiment in a laboratory with university students as their sample pool. There, the average beliefs in round one are around 64. These levels drop with learning but at a very slow rate.

controlled, the only difference between the Simultaneous parametrization A and Simultaneous B treatments is the difficulty of the problem. Thus, in line with the evidence presented in the previous section, we find that an increase in difficulty drives a wedge between observed and optimal behavior.

Result 4 (Belief Correction in Difficult Tasks)**.** *In the high-difficulty environment, delivering information through two simultaneous signals decreases the gap between the observed and Bayesian beliefs but does not eliminate it.*

4.4 Information Structure and Sequencing

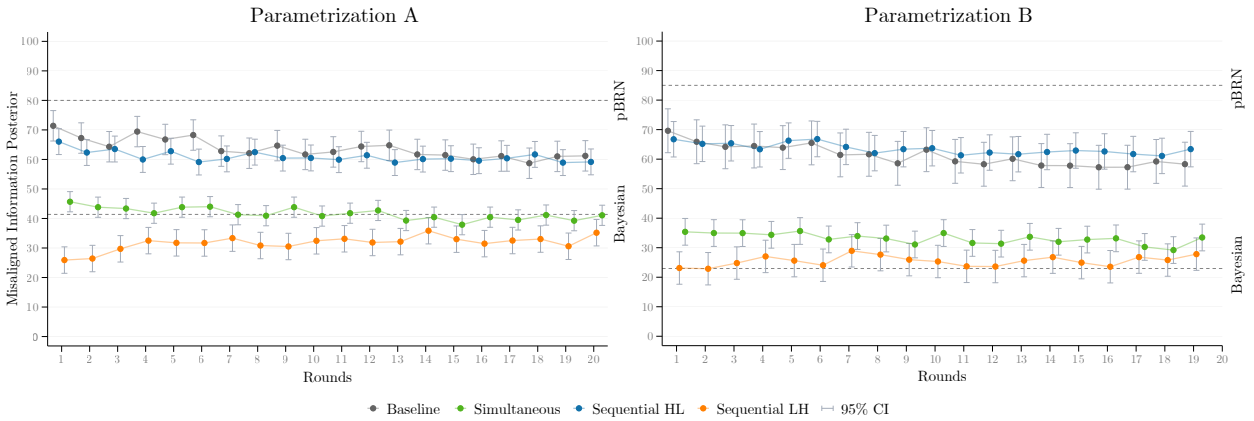
Having established the differences between the Baseline and Simultaneous treatments, we turn to examine factors that drive this difference. Two main features distinguish the Simultaneous treatment from the Baseline. First, the Baseline treatment features an information structure with an informative prior and one signal, while the Simultaneous treatment has an uninformative prior and delivers an equivalent amount of information via two signals. Second, in the Simultaneous treatment, all information is delivered at the same time, whereas the Baseline presents information sequentially; the prior precedes the signal. Either or both of these features may be responsible for the differential behavior in the two treatments.

We use our sequential information treatments to decompose the role of these features in correcting sub-optimal belief updating. Figure 7 depicts the round-by-round average posterior beliefs for all four treatments in parameterization *A*. Table 5 presents average posteriors when pooling data across all rounds.

The Effect of Information Structure per se. The informative prior in the Baseline treatment conveys an equivalent amount of information as the first signal in the Sequential HL treatment. The accuracy of the second signal is the same in both treatments, as is the sequencing of information. Thus, the difference between the two comes exclusively from the way information is communicated—the information structure.

We evaluate this comparison in two steps. First, we compare beliefs elicited after the first signal in the Sequential HL treatment with the induced prior in the Baseline treatment. While these are predicted to be identical theoretically, their empirical equivalence needs to be established. Our data directly speaks to this equivalence. Barring implementation errors, we find that participants correctly update their beliefs when faced with an uninformative prior and one signal only (see Figure 5 and the discussion in Section 4.2).

Figure 7: Posteriors in All A and B Treatments



Notes: We report the round-by-round average posteriors under misaligned information, alongside the 95% confidence intervals, clustered at the individual level. The lower horizontal dashed line depicts the Bayesian posterior, while the top dashed line depicts the perfect base-rate neglect posterior.

The second step is to compare the final beliefs elicited in the Simultaneous HL treatment with those in the Baseline. Here, as well, we find no difference in elicited beliefs at the aggregate level, see Figure 7 and Table 5. The p-value of the difference is 0.28 for parametrization A and 0.62 for parametrization B. This aggregate data, however, hides some individual differences, which we explore in detail in Section A.2 in the Appendix.[28]

Result 5 (Information Structure Effect)**.** *We document no aggregate effect on participants' reported beliefs when altering the information structure.*

The Effect of Information Sequencing. The Sequential HL and Sequential LH treatments differ only in the order in which the high- and low-accuracy signals are received. These two treatments are necessary to disentangle the effect of the order in which signals are received from the effect of, say, conditioning more heavily than theoretically optimal on the high-accuracy signal regardless of its timing.

Figure 7 shows that the sequencing of information has a substantial impact on belief updating. In all Sequential treatments, participants put a higher-than-optimal weight on the most recent signal they received. In the Sequential HL treatment with parameterization A, this means participants overweight the second (positive) signal by forming posteriors that hover around 60.89 across the twenty rounds, while in the Sequential LH treatment, they overweight the second (negative signal) and arrive at the lower-than-Bayesian

[28] To preview these results, in Section A, we show that a change in the information structure mostly affects beliefs of participants who rely on all information they receive: their beliefs in the Sequential HL treatment are less extreme compared to reported beliefs in the Baseline treatment.

posterior of 31.70. A similar pattern happens in the parameterization B, in which we observe participants overweighting the second (positive) signal by arriving at the average posterior of 63.70 in the Sequential HL treatment and the significantly smaller average posterior of 25.04 in the Sequential LH treatment.[29]

Result 6 (Recency Bias)**.** *We document a sizable recency bias independent of signal accuracy and task difficulty.*

Our results share similarities with previous literature documenting a recency bias. However, we capture this effect in a framework in which we can decompose and quantify its influence apart from other factors. Additionally, by introducing variations in the environment, we assess its robustness across parameterizations. It is worth emphasizing that we document a sizable recency bias, despite the fact that the timing between signal arrival and belief elicitation is minimal. Participants receive their first signal, their beliefs are elicited, and immediately after, posteriors incorporating the second signal are elicited. Furthermore, during this second elicitation, the interface reminds participants of the realized value of the first signal. Despite the minimal time elapsed between observing the first signal and being reminded of its realized value, we still observe a significant recency bias. If there was a substantial time gap between signal delivery, one would expect this bias to be even stronger.

Robustness Across Parameters Lastly, comparing the left and right panels of Figure 7 reveals that the ranking of estimated means across treatments remains unchanged under both parameterization A and B. The average posterior for the Baseline and the Sequential HL treatments are not statistically distinguishable, followed by a significantly lower posterior for the Simultaneous treatment and an even lower one for the Sequential LH treatment. Under both parametrizations, we find that the structure of information, i.e., the difference between delivering information via an informative prior and one signal versus an uninformative prior and two signals, has no effect on posteriors in aggregate. On the contrary, sequencing plays an important role in determining posterior beliefs. Specifically, we find a large recency bias, consistent with findings reported in Section 4.4.

Result 7 (Robustness)**.** *The ranking between all four treatments is preserved across different parameterizations, as are results regarding information structure and sequencing.*

[29] Recall from Section 4 that in our data analysis, we normalize the high-accuracy signal to be negative and the low-accuracy signal to be positive.

4.5 Countering Biases

In Section 4.4 we saw that in the low-difficulty parametrization A, the simultaneous release of information leads to elicited beliefs not statistically distinguishable from the Bayesian posterior. This is not the case in the high-difficulty parameterization B, where the simultaneous release of information does not completely align posterior beliefs with Bayesian ones, though it does bring them closer.

In the preceding sections, we highlighted two significant biases that impact belief updating. In difficult environments, when information is presented simultaneously, by failing to adequately react to the high-accuracy signal, participants tend to under-follow it. Conversely, we have also observed a pronounced recency bias, whereby participants tend to over-follow the most recent signal. With these biases in mind, we explore the possibility of mitigating the difficulty bias by releasing information sequentially, with the high-accuracy signal delivered last to utilize the sequencing bias. An examination of Figure 7 confirms that in high-difficulty environments, Sequential LH treatment yields results closer to the Bayesian posterior compared to the Simultaneous treatment. This observation is further supported by posteriors averaged across all rounds presented in Table 5: posteriors in the Sequential LH treatment are not statistically different from the Bayesian level, p-value=0.312.

Thus, the optimal information release strategy depends on the environment. In low-difficulty settings where signals are generally weighted correctly, simultaneous information release proves to be optimal. Conversely, in high-difficulty settings where there's a risk of incorrectly weighting signals, leveraging the recency bias can help mitigate this issue.

Result 8 (Countering Biases)**.** *The bias arising from sequential information arrival (recency bias) can help mitigate the difficulty bias, which arises from non-linear thinking required to reach the Bayesian posterior.*

4.6 Drivers of Base-Rate Neglect

Table 5, we present average reported beliefs for all treatments under parametrizations A ($\tilde{A}$) and B ($\tilde{B}$). We provide these estimations for both the entire dataset and for the last five rounds separately. Notably, the observed changes in means are relatively small.

Table 5: Estimated Means

	Parameters A ($\tilde{A}$)		Parameters B ($\tilde{B}$)	
	All Rounds	Last 5 Rounds	All Rounds	Last 5 Rounds
Baseline	63.79	60.43	61.93	57.97
	(1.967)	(2.423)	(2.854)	(3.447)
Simultaneous	41.65	40.29	33.39	31.77
	(0.985)	(1.293)	(1.435)	(1.678)
Sequential HL	60.89	59.95	63.70	62.35
	(1.785)	(1.966)	(2.205)	(2.538)
Sequential LH	31.70	32.56	25.04	25.79
	(1.633)	(1.849)	(2.093)	(2.464)

Individual-level clustered errors in parentheses

In both parameterizations, the Baseline treatment exhibits comparable relative levels of base-rate neglect

$$\frac{\mu^A_{Bench} - \mu^A_{Bayes}}{\mu^A_{pBRN} - \mu^A_{Bayes}} = \frac{63.79 - 41.37}{80 - 41.37} \approx 0.58, \qquad \frac{\mu^B_{Bench} - \mu^B_{Bayes}}{\mu^B_{pBRN} - \mu^B_{Bayes}} = \frac{61.93 - 22.97}{85 - 22.97} \approx 0.63.$$

In the calculations above, a score of 0 implies that, on average, participants choose the Bayesian posterior, while a score of 1 implies that, on average, participants choose the perfect base-rate neglect (pBRN) posterior. Recall that pBRN agents disregard the initial information and solely follow the signal.

Next, we delve into a decomposition of the observed level of base-rate neglect, attributing it to information structure, sequencing, and task difficulty. As discussed before, information structure has no effect on beliefs at the aggregate level, while the task difficulty and the sequencing do. The extent to which base-rate neglect is influenced by sequencing can be computed as follows

$$Sequencing^A = \frac{\mu^A_{SeqHL} - \mu^A_{Sim}}{\mu^A_{Bench} - \mu^A_{Bayes}} = \frac{60.89 - 41.65}{63.79 - 41.37} \approx 0.86.$$

$$Sequencing^B = \frac{\mu^B_{SeqHL} - \mu^B_{Sim}}{\mu^B_{Bench} - \mu^B_{Bayes}} = \frac{63.70 - 33.39}{61.93 - 22.97} \approx 0.78.$$

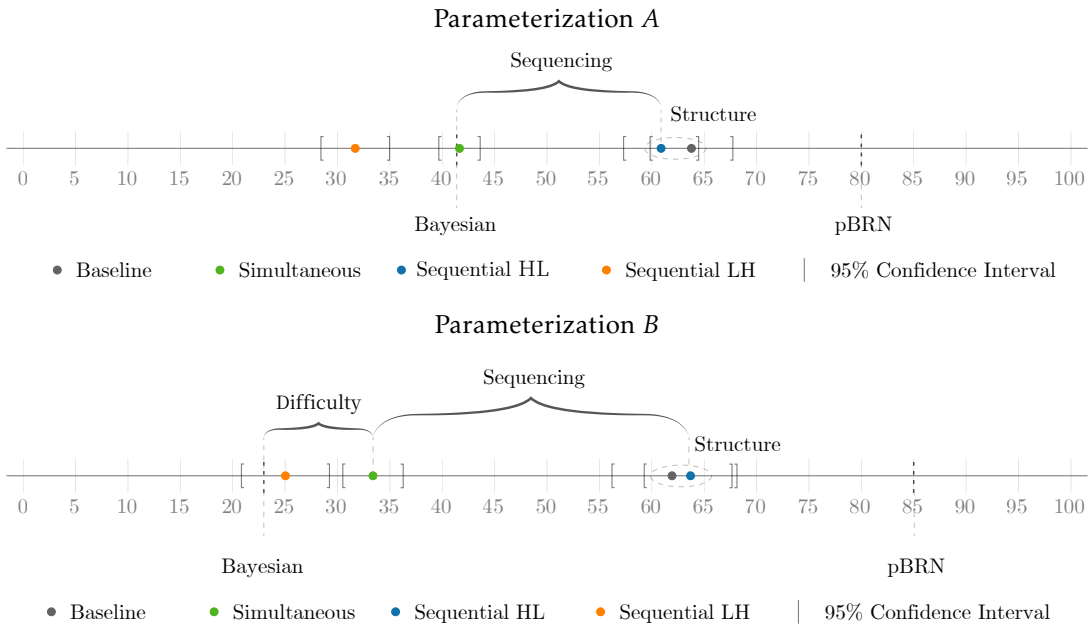
In addition, task difficulty plays a role in parameterization B and its contribution to the overall misspecified beliefs is

$$Difficulty^B = \frac{\mu^B_{Sim} - \mu^B_{Bayes}}{\mu^B_{Bench} - \mu^B_{Bayes}} = \frac{33.39 - 22.97}{61.93 - 22.97} \approx 0.27.$$

The remaining portion of base-rate neglect, although not statistically significant, is due to

information structure. We visually summarize these effects in Figure 8, which illustrates the y-axis of Figure 7 after aggregating data across rounds.

Figure 8: Average posteriors in all treatments



Result 9 (Drivers of Base-rate Neglect)**.** *Information sequencing is the main catalyst of base-rate neglect, with task difficulty also playing a significant role.*

5 Conclusions

Through a series of lab experiments, we examined how task difficulty, information structure, and information sequencing influence belief updating. Our findings highlighted that adjusting these factors can alter observed behavior from closely resembling Bayesian reasoning to exhibiting sizable deviations. Our analysis revealed that each of these elements, at the aggregate or individual level, exerts a distinct impact. We quantified these effects and explored strategies to leverage one factor against another, aiming to minimize deviations from Bayesian updating.

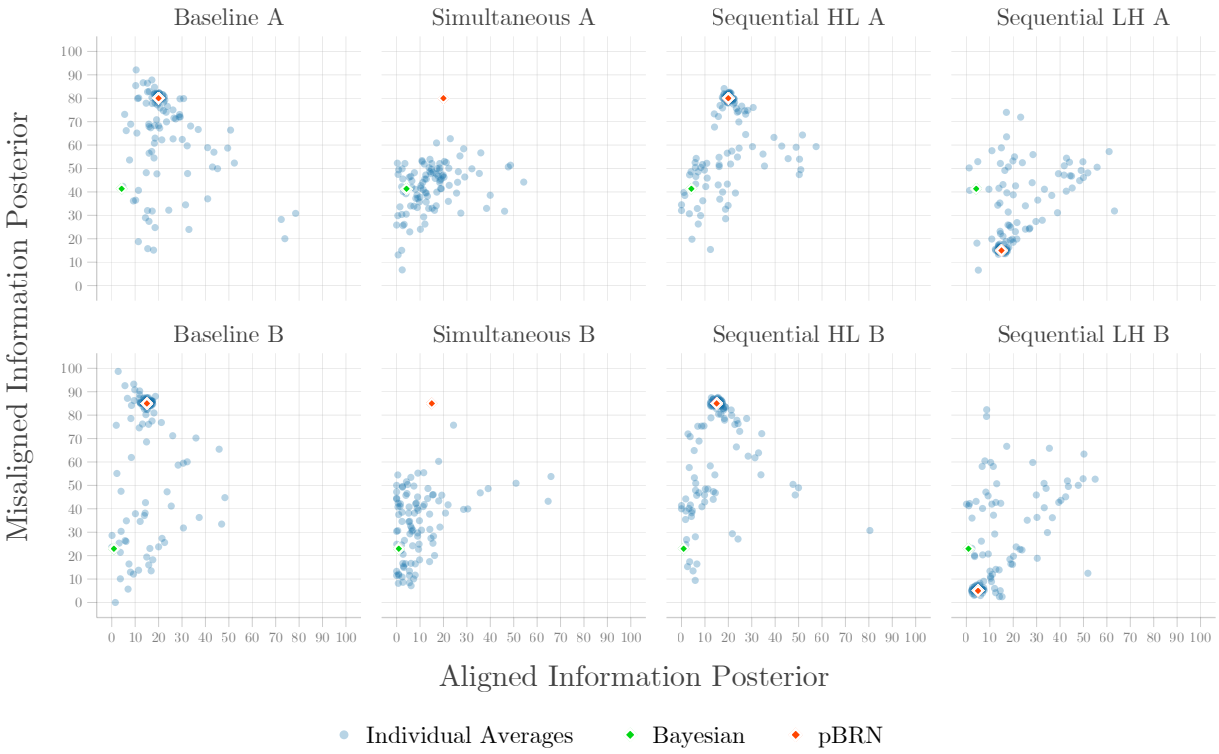
Through a range of treatments, as well as supplementary data, we conducted a comprehensive test of a notion of task difficulty rooted in the nonlinearities of the underlying problem. Taken together, this evidence suggests that nonlinearities are influential in belief updating. We believe peoples' limited ability to fully internalize nonlinearities extends beyond the realm of belief updating and presents an intriguing avenue for future research.

A Individual Level Analysis

A.1 Primary Data Patterns

We now shift our attention to individual-level behavior. For each participant, we calculate their average elicited beliefs across all rounds for both aligned and misaligned information and present these averages in Figure 9.[30] Therefore, each datapoint in the figure represents the average behavior of a single participant.

Figure 9: Average Individual Choices



Notes: To help distinguish the large amount of data bundled on the pBRN level, we apply a jitter of 1.5 magnitude. This jittering perturbs the datapoint no further than a distance of 1.5 from the initial value. The top(bottom) row displays data across treatments under parametrization A(B).

As can be seen, whenever information is released sequentially, the individual-level average posteriors are heavily bunched around the pBRN level.[31] This bunching phenomenon persists regardless of whether the sequential information delivery stems from an informative prior and a single signal (Baseline treatment) or an uninformative prior

[30] In the Online Appendix, we show the counterpart of Figure 9 utilizing only the last five rounds. All main features remain unchanged.

[31] Note that, given our normalization, in Sequential LH, the pBRN level is (15,15) under the first parametrization and (5,5) under the second.

and two signals (Sequential treatments). Only when information is released simultaneously do we observe beliefs that are not heavily concentrated around the pBRN levels. These findings align with Result 9, demonstrating the substantial impact of recency bias.

Result 10 (De-Bundling)**.** *The Simultaneous treatment is the only treatment leading to individual-level beliefs that are not strongly concentrated around the pBRN level.*

A.2 Individual-Level Effect of Information Structure

Recall that we observe no significant differences between average beliefs reported in the Baseline and the Sequential HL treatments in both parameterizations A and B (Section 4.4 and Section **??**). Although this holds true on average, in this section, we explore whether the information structure plays a role on an individual level.

In Figure 10a and Figure 10b, we show estimated kernel densities of posteriors from the Baseline and Sequential HL treatments, under parametrization A and B respectively. Within a parametrization, the estimation pools posteriors across participants and rounds.[32] In both parameterizations, despite the similar means, the distributions exhibit notable differences. In the sequential HL treatments, there is a greater concentration towards intermediate values, which are close to the Bayesian level. While the fractions of participants choosing posteriors around the pBRN levels (80 for A and 85 for B) are comparable, in the sequential HL treatment we see fewer values above these levels and fewer values for low posteriors. In both parameterizations, this mass is redistributed from the more extreme values towards the center in such a way that keeps the mean roughly unchanged. However, we run a Kolmogorov-Smirnov test between the Baseline and Sequential HL distributions and reject the null that they are the same ($p < 0.01$) under both parametrizations A and B.

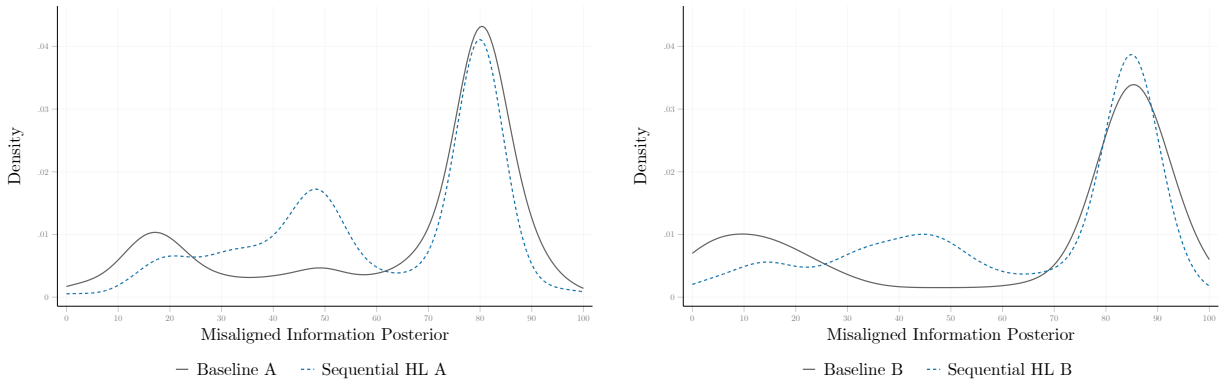
The change between the distributions can be due to small changes in the behavior of many participants, drastic changes in the behavior of some participants, or both. To further explore this, we compute the average posterior for each participant across the 20 rounds and estimate kernel densities based on these average posteriors. Doing so allows us to focus on the variation across participants. In Figure 10c and Figure 10d, we present estimated kernel densities of the average individual-level posteriors in the baseline and sequential HL treatments, under parametrization A and B respectively. As can be seen, a considerable portion of participants, on average, choose levels near the pBRN levels (80 for A and 85 for B). These participants disregard the initial information and solely follow

[32] Due to the consistent behavior exhibited by participants, conditioning on any round results in a qualitatively indistinguishable graph.

Figure 10: Distribution of Posteriors and Individual-Level Averaged Posteriors

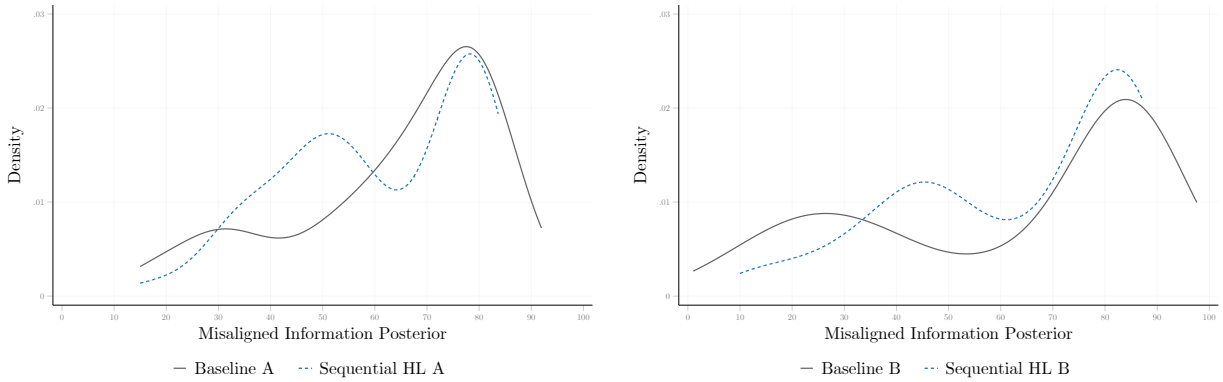
(a) Posteriors: Parameter A

(b) Posteriors: Parameter B



(c) Individual Average Posteriors: Parameter A

(d) Individual Average Posteriors: Parameter B



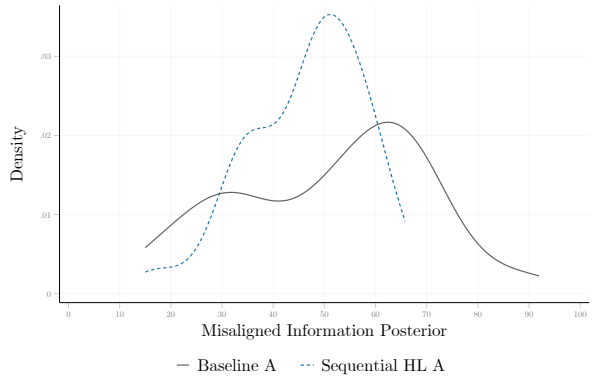
the second signal. For the remainder of the distributions, we see that in the Sequential HL treatment, fewer individuals choose extreme values. Our interpretation of these additional estimated kernel densities is that changing the information structure has no effect on individuals who solely follow the second signal, while for others, it steers their choices towards less extreme values—closer to the Bayesian level.

Below, we estimate the kernel densities under both parameterizations after removing participants that seem to behave in a pBRN manner. We remove participants from the analysis if their average aligned and misaligned posteriors are within 10 points from the pBRN posterior level.[33]

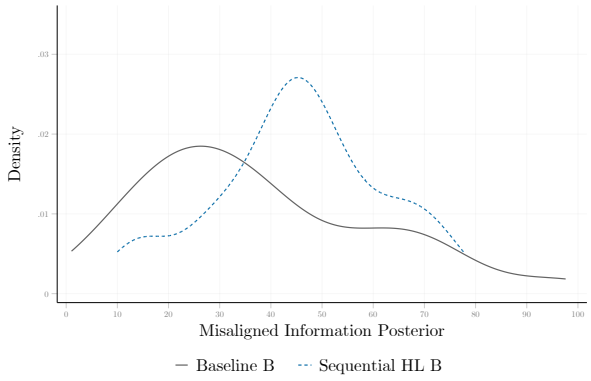
[33]For parametrization A, this implies we drop participants whose misaligned posteriors are between 70 and 90 and whose aligned posteriors are between 10 and 30. For parametrization B, this implies we drop participants whose misaligned posteriors are between 75 and 95 and whose aligned posteriors are between 5 and 25.

Figure 11: Individual-Level Averaged Posteriors Excluding pBNE

(a) Parametrization A



(b) Parametrization B



Compared to Figure 10a and Figure 10b, the difference between the estimated kernel densities becomes starker. This is in line with our interpretation that information structure seems to have minimal to no effect on participants who show pBRN behavior; however, for other non-pBRN participants, the effect is sizable. This is in line with our interpretation of the differential effect of information structure. In other words, for participants who only focus on the most recent information, the particular information structure does not play a substantial role—they ignore the initial information regardless. On the other hand, for participants who somewhat incorporate both the initial and the more recent information, the specific information structure can influence belief updating.

Result 11 (Effect of Information Structure)**.** *Elicited beliefs of participants who exclusively rely on recent information are unaffected by information structure. For other participants, a change in the information structure results in less extreme reported beliefs.*

A.3 Classifying Types: K-means Clustering

We next proceed by classifying participants into different types. To determine the number of types and the types themselves, we utilize K-means clustering, which, simply put, is a method that partitions n observations into k clusters/groups. Each observation is associated with the cluster with the nearest mean (centroid). This results in a partitioning of the data space into Voronoi cells. Specifically, K-means clustering minimizes within-cluster variances (squared Euclidean distances). This is one of the most commonly used unsupervised classifiers.[34] By employing this procedure, we bypass the need to deter-

[34] An unsupervised classifier is a machine learning algorithm that automatically identifies patterns and groups data without prior labeled training examples.

mine types subjectively. Instead, we rely on the unsupervised classification procedure to determine both the number of types and their characteristics. To determine the number of clusters, we employ two commonly used approaches, the *elbow method* and the *silhouette score*. Details of these approaches are presented in the Online Appendix. Based on this initial analysis, the suggested number of clusters is three.

Since our aim is to evaluate how the share of different types changes across treatments, we separate the typical K-means clustering into two parts. The first part, clustering, involves determining the centroids for each cluster. We do this by pooling the data across treatments within a parametrization. Having determined the centroids, we then proceed with the second part, classification, which simply associates each observation to the cluster with the nearest mean.[35] To identify the centroids, we use a standard iterative refinement technique. To summarize, we do the following: (i) determine the number of clusters, (ii) determine the centroids, and (iii) classify participants.

We follow the exercise described above for treatments one through three. We exclude the sequential LH treatment for technical reasons.[36] The clustered data is shown in Figure 12, along with the three centroids and the corresponding Voronoi sets they generate.

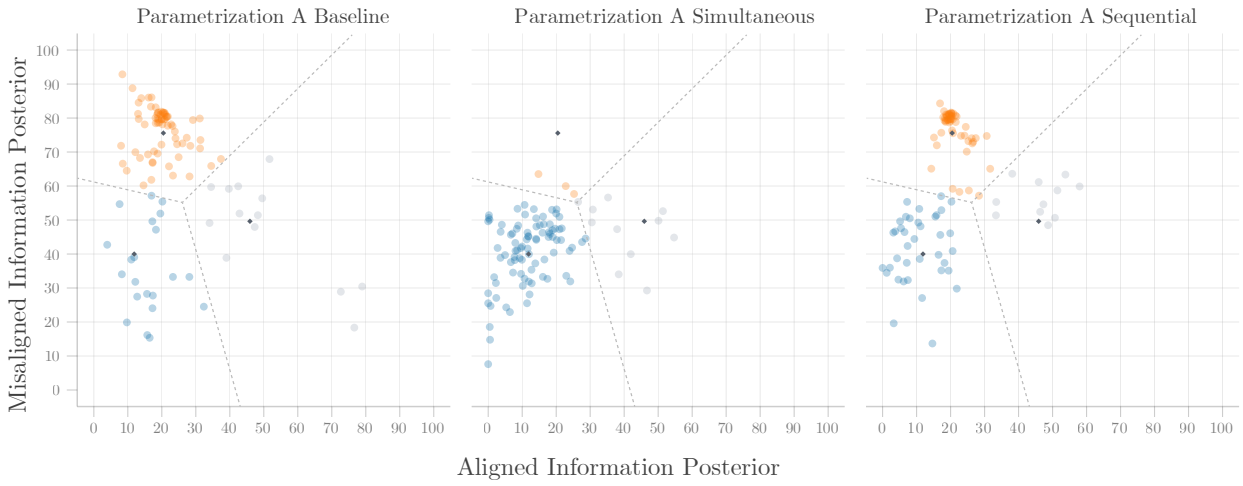
We see the emergence of three distinct clusters, with their centroids exhibiting close proximity to the Bayesian posterior (4.2,41.38), the 50-50 posterior, and the pBRN posterior (20,80). Consequently, we interpret the first group as roughly Bayesian, or closest to Bayesian, the second as potentially exhibiting confusion, and the third as roughly pBRN, or closest to pBRN. We label the second cluster as potentially confused due to the fact that, regardless of the prior and signal value, whether positive or negative, participants consistently opt for values close to 50. We display the fraction of each type across treatments in Figure 13.

As can be seen, a simultaneous release of information under the first parametrization leads to the largest share of participants classified as closest to Bayesian, with a minuscule share of agents closest to the pBRN level. Importantly, although in the previous section, we saw that the estimated mean under the Baseline and Sequential HL treatments was not statistically different, we see that the composition of the type of participants differs. The Sequential HL treatment is characterized by a higher share of closest-to-Bayesian agents

[35] Had we not followed the procedure described above, and instead, had we estimated centroids for each treatment, there would be no natural way to compare shares of participants belonging to different groups across treatments since what a group is would differ from treatment to treatment.

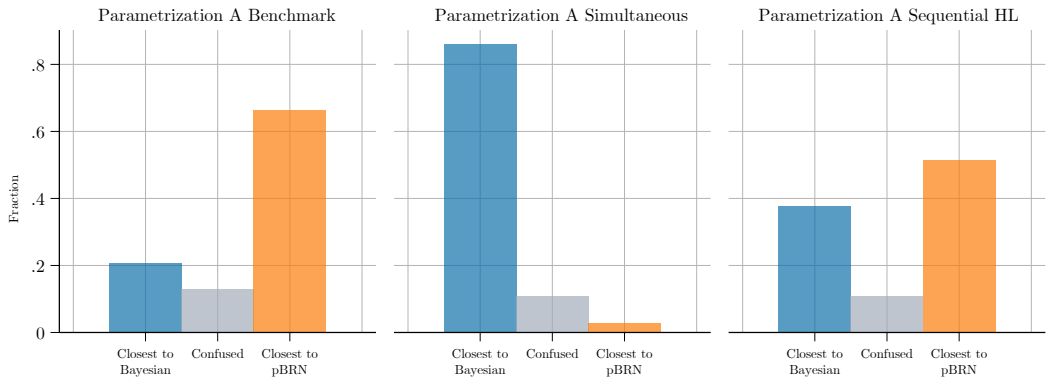
[36] In the sequential LH treatment, if we do not normalize the data, as we have done in the main analysis, the Bayesian posterior will have a different position compared to the three other treatments. If we normalize the data, the pBRN posterior will have a different position compared to the three other treatments. This, in turn, hampers our ability to have a natural interpretation of the clusters. Hence, we proceed with the clustering exercise for the first three treatments only.

Figure 12: Parametrization A Clustering



Notes: Participants are categorized into three separate clusters. Dark gray dots mark the centroids of these clusters, and dashed lines represent the Voronoi cells corresponding to these centroids.

Figure 13: Parametrization A Cluster Histogram

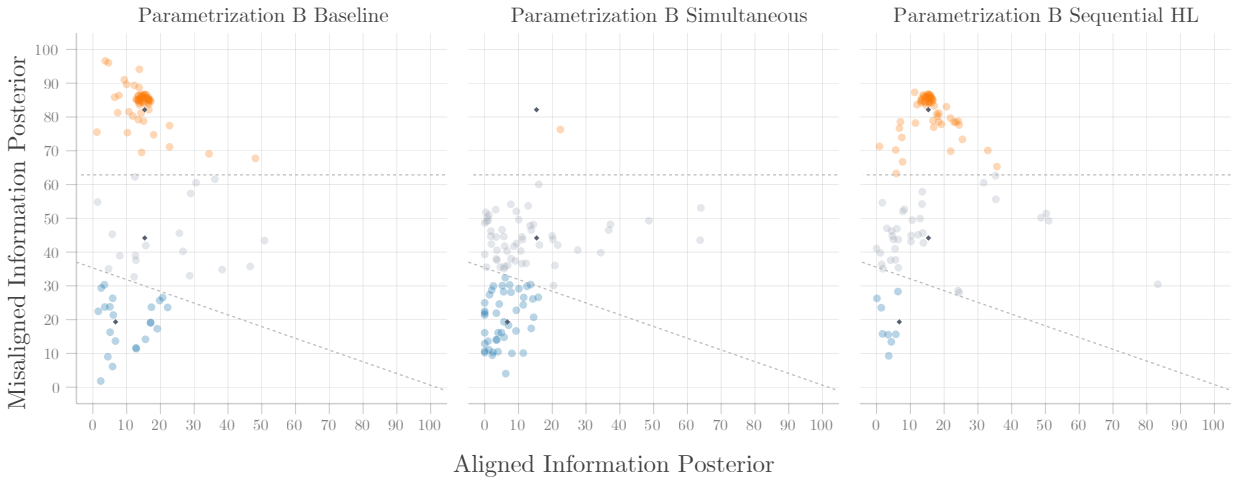


and a lower share of closest-to-pBRN agents. Thus, in line with the evidence presented in Section A.2, the information structure does seem to have an effect on individual-level behavior.

Result 12 (Information Structure and Type Classification)**.** *Information structure affects participant categorization.*

We next turn our attention to parametrization B. We once again follow the procedure described above and show the clustered data in Figure 14, along with the three centroids and the corresponding Voronoi sets they generate.

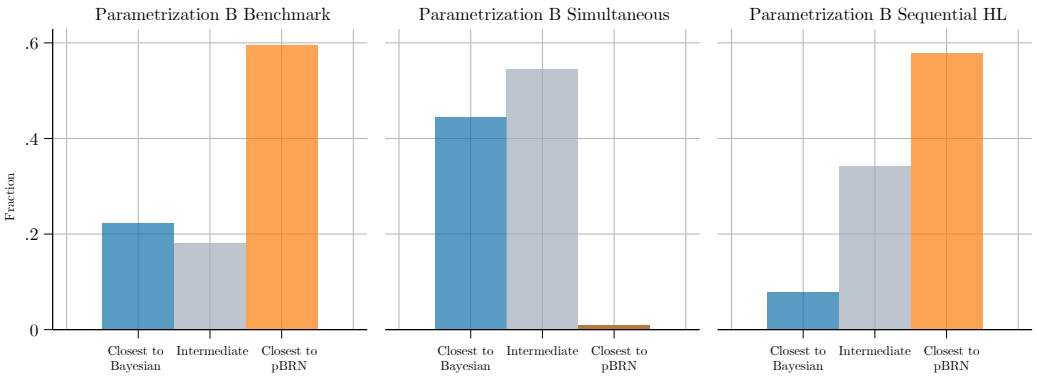
Figure 14: Parametrization B Clustering



Notes: Participants are categorized into three separate clusters. Dark gray dots mark the centroids of these clusters, and dashed lines represent the Voronoi cells corresponding to these centroids.

We see the emergence of three distinct clusters, with their centroids exhibiting close proximity to the Bayesian posterior (0.9,22.97), an in-between posterior, and the pBRN posterior (15,85). Consequently, we interpret the first group as roughly Bayesian, or closest to Bayesian, the second as in between the two extremes, and the third as roughly pBRN, or closest to pBRN. We display the fraction of each type across treatments in Figure 13.

Figure 15: Parametrization B Cluster Histogram



Once more, a simultaneous release of information leads to the largest share of participants classified as closest to Bayesian, with a minuscule share of agents classified as closest to the pBRN level. We once again see that the classification of participants differs between the baseline and sequential HL treatments.

Result 13 (Types Across Parameters)**.** *We observe variation in both clusters as well as the distribution of participants among these clusters across different parameters.*

B Concrete Specifications of Task Difficulty

B.1 Parsimonious Approach

When comparing two signals with different accuracies, the region of interest becomes the region between the juncture where both signals have the same accuracy, all the way to the point characterized by the accuracies of the signals. A Bayesian agent would be able to fully follow these changes and figure out how much more they need to weigh the high-precision signal compared to the lower-precision one. However, an agent struggling to follow such nonlinearities might make errors proportional to these cumulative nonlinearities. To link the difficulty of a task with these cumulative nonlinearities, we integrate the absolute value of the second derivative of the Bayesian posterior, starting from the juncture where two signals have equal accuracy up to the point of interest

$$C(\theta_2, \theta_1) = \int_{\theta_2}^{\theta_1} \left| \frac{d^2 P(\tilde{\theta}_1, \theta_2 | s_1, s_2)}{d\theta_1^2} \right| d\tilde{\theta}_1. \tag{3}$$

Above, $P(\tilde{\theta}_1, \theta_2 | s_1, s_2)$ represents the Bayesian posterior given signal accuracies $\tilde{\theta}_1$, θ_2, and signal realizations s_1 and s_2.[37] Low-difficulty environments will be those in which the Bayesian posterior is rather linear, and thus, neglecting nonlinearities will not matter much, resulting in low $C(\theta_2, \theta_1)$ values. In such environments, a somewhat linear approximated understanding of the environment proves effective. High-difficulty environments will be those in which the Bayesian posterior is rather non-linear. In these environments, aggregate nonlinearities are large, leading to large $C(\theta_2, \theta_1)$ values. In these environments, relying on a linear approximated understanding of the environment leads to sizable discrepancies.

For ease of exposition, we express the accuracy of a more accurate signal θ_1 in terms of the less accurate signal θ_2, i.e., $\theta_1 = \theta_2 + \eta$, where η is a positive constant.[38] The above definition of task difficulty leads to two main testable implications.

[37] This is one of many ways to capture the nonlinearities of the environment. Another measure that gives almost identical predictions in this setup is the Gini coefficient (Lorenz curve), which looks at the deviation of a graph of interest (the distribution of wealth) from the 45-degree line (a fully linear function).

[38] Naturally, the value of η is restricted to be $\eta \in (0, 1 - \theta_2)$.

1. Task Difficulty increases as the level of signal accuracies increases: $\frac{\partial C(\theta_2,\theta_2+\eta)}{\partial \theta_2} > 0$.

2. Task Difficulty increases as the gap between signal accuracies increases: $\frac{\partial C(\theta_2,\theta_2+\eta)}{\partial \eta} > 0$.

As the level and/or the gap between signal accuracies increases, the updating task takes place on a more nonlinear region, which, according to our predictions, may lead participants to make larger mistakes.[39]

B.2 Modified Grether Model

Grether (1980) proposed a generalization of Bayesian updating that accommodated a variety of deviations commonly observed in experiments. This generalization is extensively used in the literature on beliefs Benjamin (2019). In this section, we explore a modification of the Grether (1980) model that yields comparable qualitative predictions to our task difficulty concept discussed in Section 2.3.

Grether (1980) writes the posterior-odds ratio in the following form

$$\frac{\pi(S|s_2,s_1)}{\pi(F|s_2,s_1)} = \left(\frac{P(s_2|S)}{P(s_2|F)}\frac{P(s_1|S)}{P(s_1|F)}\right)^{\alpha}\left(\frac{P(S)}{P(F)}\right)^{\beta}$$

where $\pi(\cdot)$ captures an agent's possibly biased beliefs. If $\alpha = \beta = 1$, the model reduces to Bayesian updating, whereas $\alpha < 1$ ($\beta < 1$) implies underinference, extracting less information from the signal (prior) than prescribed by Bayes' rule. For comparability purposes, we focus on the uninformative prior case with $\frac{P(S)}{P(F)} = 1$.

We make two modifications to this formulation, given an uninformative prior: (i) agents properly update beliefs when receiving only one signal, and (ii) in the presence of more than one signal, agents under-follow signals more, the more accurate signals are. The first modification relates to the idea that task difficulty, or high nonlinearities, only appear when there is more than one source of information. The second modification captures the idea that under Bayesian updating, agents are expected to react more and more to small changes in signals' accuracies as these accuracies grow larger. It is in such cases that, we believe, the aforementioned sluggishness is emphasized, and individuals fail to properly Bayesian update.

[39] When both signals have the same realizations, there are parameter values that make the effect of the level non-monotonic. However, our study focuses on cases with misaligned information, as declared in the preregistration and described further in Section 3. For misaligned signals, the above predictions hold true for all parameter values.

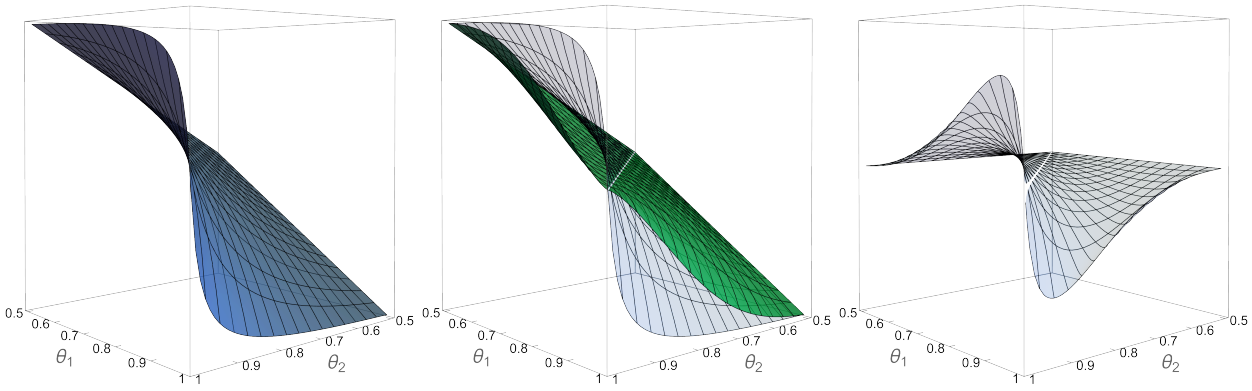
One way to accommodate these two modifications in the Grether (1980) model is to replace parameter α with a function $\gamma(\theta_1, \theta_2)$ that depends on the signals' accuracies:

$$\gamma(\theta_1,\theta_2) = \alpha \cdot \underbrace{(2min(\theta_1,\theta_2)-1)}_{\in[0,1],\text{ weight on }\alpha} + 1 \cdot \underbrace{(2-2min(\theta_1,\theta_2))}_{\in[0,1],\text{ weight on }1} \qquad 0 \leq \alpha \leq 1 \qquad (4)$$

Note that the above functional form reflects the two proposed modifications. Indeed, if $min(\theta_1, \theta_2) = 1/2$ then $\gamma(\theta_1, \theta_2) = 1$ in line with the first modification and as $min(\theta_1, \theta_2)$ increases $\gamma(\theta_1, \theta_2)$ decreases, in line with the second modification.

To illustrate, Figure 16 focuses on the case of two misaligned signals and compares the Bayesian posterior with the one induced by the modified Grether model across different θ_1 and θ_2 values.[40] By design, the functions are in agreement when one signal is uninformative (θ_1 or θ_2 is 0.5). The functions are also in agreement if both signals have equal accuracy ($\theta_1 = \theta_2$) because, in such cases, both models assign equal weight to each signal.[41]

Figure 16: Gap Between Bayesian Posterior and Modified Greather



Notes: The figures above show the Bayesian posterior (left), a transparent Bayesian posterior, and the posterior of the modified Grether model (middle), as well as their difference (right). The graphs are plotted for values θ_1 and $\theta_2 \in [0.5, 0.99]$. The value of α is set to 0.

In all other cases, the two functions differ. If signal accuracies are low or if the gap

[40] The posterior belief implied by the modified Grether model for two misaligned signals is

$$\pi(S|s_2 = p, s_1 = n) = \frac{\left(\frac{\theta_2}{1-\theta_2}\frac{1-\theta_1}{\theta_1}\right)^{\gamma(\theta_1,\theta_2)}}{\left(\frac{\theta_2}{1-\theta_2}\frac{1-\theta_1}{\theta_1}\right)^{\gamma(\theta_1,\theta_2)} + 1}$$

[41] The two functions are also in agreement if one of the signals is fully informative (θ_1 or θ_2 is 1). To make displaying the graphs clearer, this region is clipped off. Figure 16 only displays posteriors for θ_1 and $\theta_2 \in [0.5, 0.99]$.

between the signal accuracies is small, the modified Grether model results in posteriors close to Bayesian ones. However, when signals have very different accuracies, or when both signals are very accurate, the modified Grether model produces noticeably more inert updating than the Bayesian one. This is the region in which the sluggishness of our model plays a large role, leading to large differences. These predictions mimic those described in Section 2.3.

B.3 Linear Thinking Model

In this section, we provide an alternative model that captures the difficulty of incorporating nonlinearities in belief-updating tasks. We assume that agents struggle to fully appreciate rapid changes required to form correct posterior beliefs in response to small changes in fundamentals and, instead, are only able to adapt to these changes partially. Focusing on the case of misaligned signals, let $\tilde{\pi}(S|s_2 = p, s_1 = n)$ be the posterior of an agent who struggles to incorporate nonlinearities. We define the derivative of this posterior to be a weighted average with weight $\alpha \in [0,1]$ on the derivative of the Bayesian posterior and a constant. A constant is chosen to respect the idea that agents update beliefs correctly when there is only one signal, i.e.,

$$\frac{d\tilde{\pi}(S|s_2 = p, s_1 = n)}{d\theta_1} = \alpha \frac{dP(S|s_2 = p, s_1 = n)}{d\theta_1} + (1-\alpha) \overbrace{\frac{dP(S|s_2 = p, s_1 = n)}{d\theta_1}\Big|_{\theta_2 = \frac{1}{2}}}^{\text{a constant}}$$

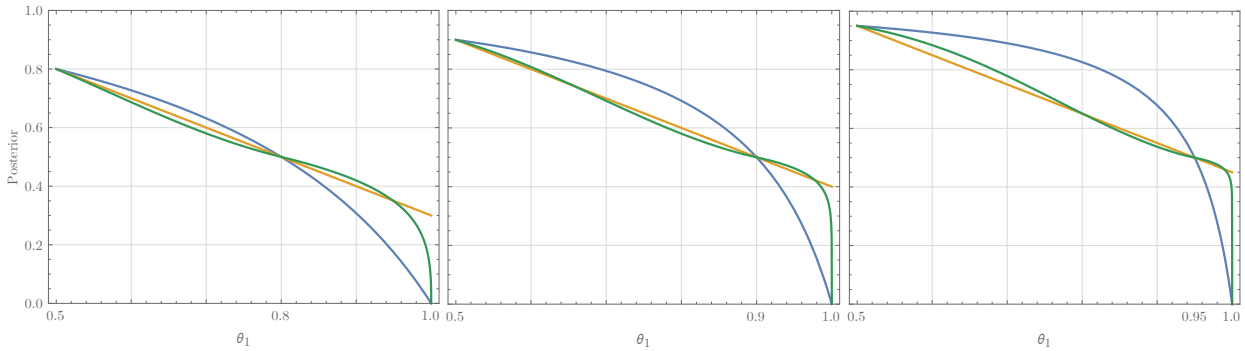
The derivative with respect to the other signal accuracy is similar. Solving the above partial differential equation and using the fact that when $\alpha = 1$ the equation reduces to the Bayesian posterior, yields

$$\tilde{\pi}(S|s_2 = p, s_1 = n) = \alpha \underbrace{\frac{\theta_2(1-\theta_1)}{\theta_2 + \theta_1 - 2\theta_2\theta_1}}_{\text{Bayesian Posterior}} + (1-\alpha) \underbrace{\left(\frac{1}{2} - \theta_1 + \theta_2\right)}_{\text{Fully Linear}}$$

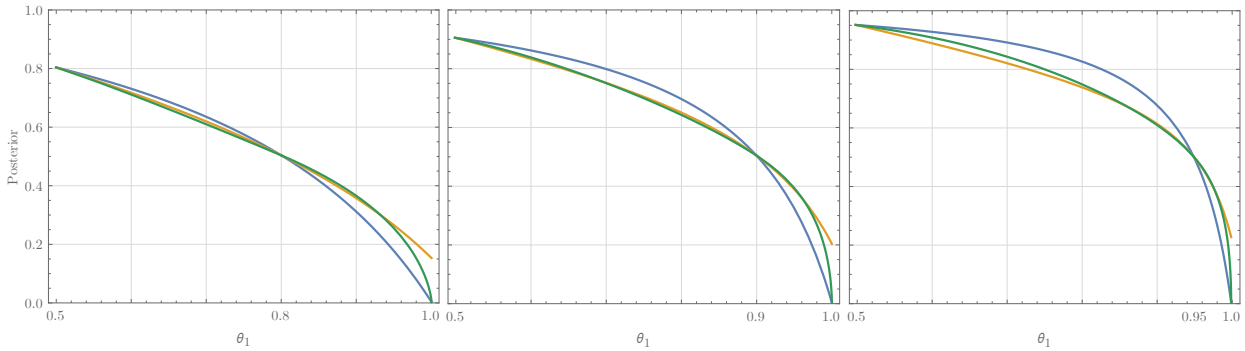
Note that since the first derivative is a convex combination of the Bayesian first derivative and a constant, the second derivative will always be lower in magnitude than the Bayesian second derivative.

As a final step, we compare the Bayesian posterior $P(S|s_2 = p, s_1 = n)$ with the posterior from the modified Grether model $\pi(S|s_2 = p, s_1 = n)$ discussed in Section B.2, and the linear thinking posterior derived in this section $\tilde{\pi}(S|s_2 = p, s_1 = n)$. We show these posteriors for various levels of θ_1 with $\alpha = 0$ in the first row and $\alpha = 1/2$ in the second row of Figure

Figure 17: Model Prediction Comparison



Notes: Above, from left to right, we utilize $\theta_2 = \{0.80, 0.90, 0.95\}$ and $\alpha = 0$. The blue, green, and orange graphs represent the Bayesian, modified Grether, and Linear Thinking posteriors, respectively.

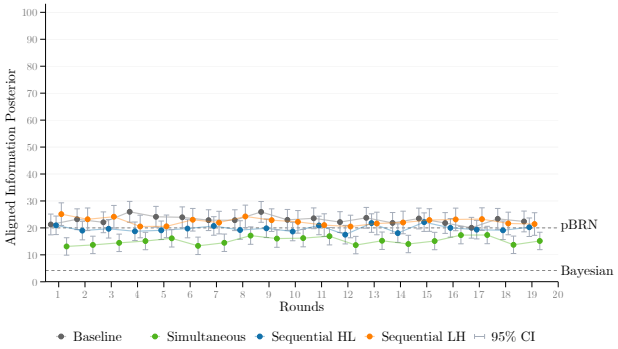


Notes: Above, from left to right, we utilize $\theta_2 = \{0.80, 0.90, 0.95\}$ and $\alpha = 1/2$.

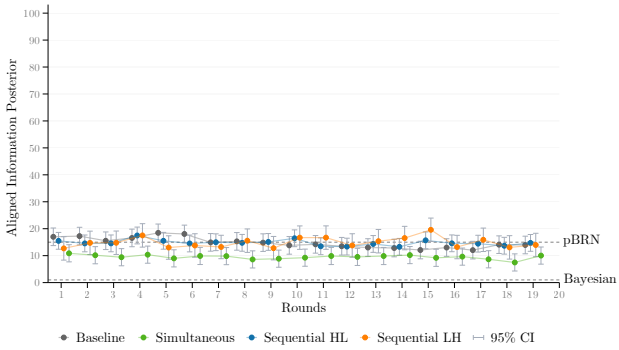
17. The modified Grether and the linear thinking model produce similar predictions, both in the direction of departure from Bayesian updating as well as in magnitude. The main difference between the two models emerges when one of the signals becomes fully informative ($\theta_1 = 1$). Near this region, the modified Grether model quickly converges to 0, while the linear thinking model does not. Again, since the fully informative case is not the focus of this study, for relevant regions (away from $\theta_1 = 1$), for any value of α, the two models produce almost identical predictions. Importantly, under both models, when the level of signal accuracies is low, the departure from Bayesian updating is small, even for sizable differences in signal accuracies. In contrast, when the level of signal accuracies is high, even small differences in signal accuracies lead to large gaps from the predicted Bayesian behavior.

Figure 18: Aligned Information Posteriors

(a) Parametrization A



(b) Parametrization B



C Additional Analysis

C.1 Aligned Information Posteriors

In Figure 18a and Figure 18b, we graph the average round-by-round beliefs of participants when information is aligned. For the Baseline treatment, these are cases when the realized signal is in the direction in which the prior leans. For all other treatments, these are cases in which both signals have the same realized value.

C.2 Task Difficulty and Cognitive Uncertainty: Additional Calculations

The sequential arrival of information is less than ideal for the study of task difficulty, as sequencing also impacts elicited beliefs. However, utilizing symmetric parameter cases, we can perform a back-of-the-envelope calculation to account for the effect of sequencing. Consider the (70,70) and (90,90) treatments, where the first (second) number corresponds to the accuracy of the prior (signal). Given our measure of task difficulty, these are both low-difficulty treatments. The Bayesian posterior is 50 for both. Data reveals that the mean posteriors are 56 and 53, respectively. Thus, $gap_{low} \in [3,6]$, which is rather small despite the fact that belief elicitation is hindered by sequencing. Now, consider two examples with relatively high difficulty (90,70) and (70,90). Because we, once again, normalize the high accuracy signal/prior to be negative, the Bayesian posterior in both cases is 20.6. The data reveals that in the former case, the posterior is 43, whereas in the latter case, the posterior is 37. This gap (between 43 and 37) is due to sequencing; in the former case, the negative and high-accuracy information arrives first, whereas, in the

second case, it arrives later. Naturally, a treatment in which all information arrives simultaneously must lie between these two values. Consequently, the gap, in this case, must be $gap_{high} \in [17.6, 23.6]$. Thus, the gap has risen from $[3, 6]$ to $[17, 23.6]$, implying that the gap has increased by at least 11.23 or at most 20.6 percentage points. In the former (later) case, task difficulty would account for $^{11.23}/_{17} = 66\%$ ($^{20.6}/_{23.6} = 87\%$) of the gap. Hence, while indirectly accounting for the effect of sequencing, we see that the bulk of the increase in the gap can not be related to sequencing and, therefore, must be difficulty related.

References

Augenblick, N., Lazarus, E., and Thaler, M. (2023). Overinference from weak signals and underinference from strong signals. *working paper*.

Ba, C., Bohren, A., and Imas, A. (2023). Over- and underreaction to information. *working paper*.

Banovetz, J. and Oprea, R. (2022). Complexity and procedural choice. *manuscript*.

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

Barbey, A. and Sloman, S. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30.

Becker, G., DeGroot, M., and Marschak, J. (1964). Measuring utility by a singleresponse sequential method. *Behavioral Science*, 9:226–232.

Benjamin, D., Bodoh-Creed†and, A., and Rabin, M. (2019). Base-rate neglect: Foundations and implications. *working paper*.

Benjamin, D., Rabin, M., and Raymond, C. (2016). A model of nonbelief in the law of large numbers. *Journal of European Economic Association*, 14(2):515–544.

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

Bernheim, B. D. and Sprenger, C. (2020). On the empirical validity of cumulative prospect theory: experimental evidence of rank-independent probability weighting. *Econometrica*, 88(4):1363–1409.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14:375–398.

Camara, M. (2021). Computationally tractable choice. *manuscript*.

Camerer, C. (1987). Do biases in probability judgment matter in markets? experimental evidence. *The American Economic Review*, 77(5):981–997.

Charness, G. and Levine, D. (2005). When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect. *American Economic Review*, 95(4).

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Cheng, P. and Holyoak, K. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17(4):391–416.

Danz, D., Vesterlund, L., and Wilson, A. (2021). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883.

Edenborough, R. (1975). Order effects and display persistence in probabilistic opinion revision. *Bulletin of the Psychonomic Society*, 5 (1):39–40.

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

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics*.

Enke, B., Graeber, T., and Oprea, R. (2023). Complexity and time. *manuscript*.

Enke, B. and Shubatt, C. (2023). Quantifying lottery choice complexity. *working paper*.

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

Esponda, I., Vespa, E., and Yuksel, S. (2023). Mental models and learning: The case of base-rate neglect. *American Economic Review*.

Fan, T., Liang, Y., and Peng, C. (2022). The inference-forecast gap in belief updating. *manuscript*.

Fudenberg, D. and Puri, I. (2022). Evaluating and extending theories of choice under risk. *manuscript*.

Ganguly, A., Kagel, J., and Moser, D. (2000). Do asset market prices reflect traders' judgment biases? *Journal of Risk and Uncertainty*, 20(3):219–245.

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

Gneezy, U., Enke, B., Hall, B., Martin, D., Nelidov, V., Offerman, T., and van de Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics Studies*.

Gneezy, U., Rockenbach, B., and Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior and Organization*, 93:293–300.

Graeber, T. (2023). Inattentive inference. *Journal of the European Economic Association*, 21(2):560–592.

Grether, D. (1992). Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1).

Grether, D. M. (1978). Recent psychological studies of behavior under uncertainty. *American Economic Review*.

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

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

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

Koehler, J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenge. *Behavioral and Brain Sciences*, 19.

Levy, M. and Tasoff, J. (2016). Exponential-growth bias and lifecycle consumption get access arrow. *Journal of the European Economic Association*, 14(3):545–583.

Levy, M. and Tasoff, J. (2017). Exponential growth bias and overconfidence. *Journal of Economic Psychology*, 58:1–14.

Madrian, B. C. and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, 116(4):1149–1187.

Oprea, R. (2022). Simplicity equivalents. *manuscript*.

Oprea, R. D. (2020). What makes a rule complex. *American Economic Review*, 110(12):3913–3951.

Pitz, G. and Reinhold, H. (1968). Payoff effects in sequential decision-making. *Journal of Experimental Psychology*, 77 (2):249–257.

Puri, I. (2022). Preference for simplicity. *manuscript*.

Rees-Jones, A. and Taubinsky, D. (2020). Measuring schmeduling. *Review of Economic Studies*, 87:2399–2438.

Stango, V. and Zinman, J. (2009). Exponential growth bias and household finance. *Journal of Finance*, 64(6):2807–2849.

Thaler, R. H. and Benartzi, S. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, 112(51):5164–5187.

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press: New Haven & London.

Valone, T. J. (2006). Are animals capable of bayesian updating? an empirical review. *Oikos*, 112:252–259.

Wagenaar, W. and Sagaria, S. (1975). Misperception of exponential growth. *Perception and Psychophysics*, 18(6):416–422.