Weighting Competing Models

Chiara Aina* Florian H. Schneider†

January 26, 2026

Abstract

We study how individuals update their beliefs in the presence of competing data-generating processes, or models, that could explain observed data. Through experiments, we identify the weights participants assign to different models and find that the most common updating rule gives full weight to the model that best fits the data. While some participants assign positive weights to multiple models—consistent with Bayesian updating—they often do so in a systematically biased manner. Moreover, these biases in model weighting frequently lead participants to become more certain about a state regardless of the data, violating a core property of Bayesian updating.

Keywords: Belief Updating, Narratives, Mental Models, Experiments
JEL classification: D83, D9, C90

*Universitat Pompeu Fabra and Barcelona School of Economics. E-mail: chiara.aina@upf.edu.
†Department of Economics and Center for Economic Behavior and Inequality (CEBI), University of Copenhagen. E-mail: flsc@econ.ku.dk.

We are grateful to Sandro Ambühl, Andrea Amelio, Kai Barron, Aislinn Bohren, Katharina Brütt, Pol Campos-Mercade, John Conlon, Agata Farina, Mira Frick, Alex Imas, Alessandro Ispano, Sevgi Yuksel, Giacomo Lanzani, Nick Netzer, Joshua Schwartzstein, Marta Serra-Garcia, Andrei Shleifer, Peter Norman Sørensen, Jakub Steiner, Adi Sunderam, Michael Thaler, Roberto Weber, Florian Zimmermann, as well as the participants at UAB, Capri In Theory, CEU, CUNEF, Columbia, DICE, IZA Beliefs Workshop, Memory and Attention Conference, MIT Sloan, UCPH, UPF, and UZH. Chiara Aina acknowledges financial support from the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2024-001476-S), funded by MCIN/AEI/10.13039/501100011033, and JDC2023-052677-I financed by MCIU/AEI/10.13039/501100011033 and by the ESF+. Florian H. Schneider acknowledges funding from the University of Copenhagen and the Danish National Research Foundation Grant DNRF134. Our studies obtained ethical approval from the Research Ethics Committee at the Department of Economics at the University of Copenhagen and the Committee on the Use of Human Subjects at Harvard University.

1 Introduction

Economic decision-making often occurs in situations where individuals have to interpret new information to form expectations about relevant outcomes. A large body of literature investigates how people revise beliefs in response to new information, focusing on situations where a single data-generating process, or *model*, provides a clear interpretation of the data. However, in many contexts, individuals are faced with multiple, conflicting interpretations, making it unclear how the data have been generated and how they should be interpreted.

For example, in asset markets, investors often use data on past performance to revise beliefs about future price trends but may encounter conflicting models to predict future returns based on historical returns. Such models might include mean reversion—"the asset's price went up in the past, so it will go down in the future"—, momentum—"the asset's price went up in the past, so it will continue to rise"—, or random walk—"past price movement provides no information about future performance" (e.g., Barberis et al., 1998). Similarly, in politics, contrasting views about election integrity, such as "vote counting is unbiased" versus "elections are rigged," lead to different beliefs about the legitimate winner of an election based on the same reported election result. In healthcare, debates about the safety of medical interventions arise because of contrasting views on whether "vaccine tests are trustworthy" or "vaccine tests are unreliable." Despite the prevalence of settings where people encounter conflicting models, little is known about how they update their beliefs in such situations.

This paper uses experiments to study how individuals update their beliefs when confronted with conflicting models that could explain the observed data. As in the examples above, we focus on a setting where participants aim to learn about an unknown state of the world—such as future asset returns, the legitimate winner of an election, or vaccine safety—based on a potentially informative yet noisy signal. Models provide distinct interpretations of the signal, proposing a data-generating process that quantifies the link between the signals and the states. Participants are presented with multiple models that may have generated the signal and make opposing predictions about the true state given the observed signal. We then analyze how individuals weight the models to aggregate the corresponding predictions about the true state. Specifically, we study which potentially biased updating rules—introduced below—describe how they weight models.

Our findings offer an empirical basis for a rapidly growing body of theoretical literature that formalizes narratives as models, as introduced by Schwartzstein and Sunderam (2021). Papers within this literature assume *model selection*, where agents form predictions about the state by assigning full weight to a single model. This relates to the tendency to simplify Bayesian reasoning into categorical thinking (Mullainathan, 2002; Mullainathan et al., 2008). Among plausible criteria for model selection, the most common assumption is that agents adopt the model that best fits the data, which we refer to as *model selection via maximum likelihood*. This approach is also considered in the literature

on ambiguity aversion (e.g., Dempster, 1967; Shafer, 1976; Gilboa and Schmeidler, 1993). Notably, placing full weight on a single model contrasts sharply with Bayesian updating: to revise their beliefs, a Bayesian decision maker retains multiple model predictions and appropriately weights them by their likelihood of being the true data-generating process given the observed signal. Alternatively, agents might assign positive weight to multiple models but in a biased manner. Our experiment allows us to differentiate between these various rules and examine their implications for belief updating.

Studying belief updating with naturally occurring data poses the key challenge that the models individuals might consider are unknown, making it difficult to distinguish between the predictions of different updating rules. To overcome this challenge, we design a novel experimental paradigm that allows us to construct and control the models that may have generated the data. Our approach builds on the classic balls-and-urns updating task (Edwards, 1968) and closely aligns with the theoretical papers that formalize narratives as models. Participants observe a signal and are tasked with forming accurate beliefs about an unknown state. Unlike the classical task where participants encounter only one model, our experiment presents participants with two models, one of which is randomly selected as the true data-generating process responsible for producing the observed signal. We elicit participants' beliefs about the state in a series of seven of these updating tasks featuring different model pairs to test the robustness of our results and assess participants' consistency in applying updating rules.

To investigate biases in model weighting, it is crucial to distinguish these from other biases in belief updating, especially those related to deriving the model predictions from the observed signal (Bohren and Hauser, 2024). We address this by providing participants with the Bayesian prediction for each model upon observing the signal. This approach eliminates biases in deriving model predictions—well-documented in the literature—enabling us to identify model weights and classify reported beliefs according to different updating rules. Providing the model predictions mirrors a realistic feature of many real-world settings where individuals encounter a model alongside its prediction. For example, in asset markets, a financial advisor might advise an investor to buy an asset that has recently risen in price, explaining the concept of "momentum," along with the prediction that the asset's price will continue to rise in the future. Or, a politician who lost an election might tell voters that the "elections were rigged," and thus, that they are the rightful winner.

We find that the most frequent updating rule in our data is selecting the best-fitting model, which accounts for one-third of all elicited beliefs. This result provides strong evidence for model selection via maximum likelihood. Another twelve percent of beliefs correspond to selecting the worst-fitting model. What could lead participants to select the worst-fitting model? This could be attributed to occasional errors in identifying the best-fitting model, resulting in a stochastic version of model selection via maximum likelihood, or to the use of alternative criteria for model selection. We explore two plausible alternative criteria: participants might choose models either based on their *informa-*

tiveness given the observed signal (inspired by Yang, 2023) or based on characteristics unrelated to the signal, which we refer to as *dogmatic model selection*, a concept related to common assumptions in the literature on learning with misspecified models (see the discussion in Ba, 2024). Unlike the maximum likelihood criterion, these alternatives have limited predictive power. Instead, our findings predominantly support a stochastic version of model selection via maximum likelihood. This is further confirmed by a simple calibration exercise: model selection in our data is well-described by a stochastic version of model selection via maximum likelihood, with only a small minority of participants employing other model selection criteria.

Model selection is a widespread phenomenon in our data: nearly half of participants' beliefs are consistent with using only a single model to update beliefs. However, a substantial portion of participants assign positive weights to both models, albeit in a biased way. Only a small percentage of guesses (8%) align with Bayesian updating. Instead, a larger share (18%) is consistent with an extreme form of underinference about the models, where participants weight the model predictions by the prior rather than by the posterior over the models. We refer to this as *one-stage updating*: participants recognize that the signal is informative for learning about the state given each model but fail to recognize that the signal is also informative about the underlying model, resulting in an improper combination of the model predictions. Such an updating rule is assumed in the literature on persuasion with coarse thinking (Mullainathan et al., 2008). One-stage updating is the second most common rule in our data, following model selection via maximum likelihood, which represents an extreme form of overinference about the models.

Notably, model selection, Bayesian updating, and one-stage updating can account for over 70% of all elicited beliefs. Hence, these few rules capture the participants' belief updating to a large extent. We also explore alternative updating rules, and find that some other participants employ less extreme forms of over- or underinference about the models. Our analysis finds no evidence that any other rules are applied frequently or consistently in our data. A direct implication of the frequent use of these different updating rules is that the distribution of beliefs about the state of the world is multimodal, as in Bordalo et al. (2026), with peaks at the predictions of the best-fitting model, the worst-fitting model, and the one-stage updating.

Participants show a high degree of consistency in applying these updating rules. We show this with two approaches. First, we design the study such that each participant encounters some model pairs twice; when participants observe different signal realizations in these tasks, we can examine the posterior beliefs reported by participants for both signals, which we refer to as *vectors of posteriors*. The most frequently reported vector of posteriors corresponds to selecting the best-fitting model for both signal realizations, followed by the one predicted by one-stage updating. Second, we investigate whether participants follow the same updating rules in all seven updating tasks. We find that more than 40% of the participants use the same rule for all seven tasks, 49% when allowing them to

deviate from a rule in one task, and 60% when requiring only a majority of consistent guesses. Thus, most participants use one of the updating rules consistently throughout the experiment.

Importantly, the data on the reported vectors of posteriors also allow us to test a key prediction from the literature on model persuasion: when presented with conflicting models, individuals may become more certain about a state regardless of the observed data. We refer to such beliefs as *Bayes-inconsistent*, as they violate a basic property of Bayesian updating. When participants encounter conflicting models and consistently select the best-fitting model, they are predicted to hold Bayes-inconsistent beliefs (Aina, 2025). Our data reveal that half of the reported vectors of posteriors are Bayes-inconsistent, where the reported guesses for both signals are both either higher or lower than the prior. This widespread emergence of Bayes-inconsistencies underscores both the relevance and the power of persuasion using conflicting models and contrasts with classic persuasion theories, such as Bayesian persuasion (Kamenica and Gentzkow, 2011). Moreover, these inconsistencies exhibit a pattern aligned with confirmation bias, which is the tendency to confirm one's beliefs (e.g., Rabin and Schrag, 1999; Fryer et al., 2019).[1] While in previous studies this pattern can be attributed to motivated beliefs or strong initial beliefs, we find that Bayes-inconsistent beliefs can also arise in neutral settings—where individuals have neither a personal stake nor a preexisting stance—due to the presence of conflicting models and biases in model weighting.

To study the generalizability of our findings, we conduct additional studies where we independently vary features of the updating environment and examine whether model selection remains prevalent. First, we examine a setting where participants are not given an objective prior over the models. This allows us to provide insights into situations where individuals encounter multiple models without an objective distribution over them, a setting that is also relevant to the literature on ambiguous beliefs. Second, we investigate a setting where model predictions are not immediately available, requiring participants to exert some effort to reveal each of them. This endogenous acquisition of model predictions helps us shed light on settings where these predictions are not readily available, and individuals need to take deliberate action to access them, reflecting search costs, cognitive effort, or attentional constraints. Our data show that almost all participants deliberately seek model predictions. Third, we consider a setting where participants face non-conflicting models, that is, models that make different predictions in magnitude but agree on the direction of the update. This comparison allows us to study whether participants are more willing to weight two non-conflicting models than fundamentally opposing ones. Finally, we study a setting with more than two models by adding a third model. In

[1]Lord et al. (1979) and subsequent studies instructed participants to review information on controversial topics, assess whether it supported or opposed the issue, and report any shifts in their beliefs. These studies consistently found that participants' posterior attitudes became more favorable if they were initially supportive or more opposed if they were initially critical. This suggests that individuals may adjust their beliefs in a preferred direction, regardless of the evidence, exhibiting a pattern closely related to Bayes-inconsistent beliefs.

particular, in one of the tasks, the third model corresponds to the Bayesian predictions. This allows us to test the prevalence of model selection via maximum likelihood when the Bayesian prediction is as salient as the other model predictions and to study the persuasiveness of a model that provides the correct Bayesian prediction.

Our data shows that the updating patterns documented in our main study extend to all these environments. Compared to a baseline condition that replicates our main study, neither treatment leads to meaningful changes in updating behavior. Model selection remains the prevalent approach in all settings. Consistent with our main finding, selecting the best-fitting model is the most frequent updating rule across all treatments, even when the Bayesian prediction is readily available and as salient as the best-fitting one.

In summary, we find evidence supporting the theoretical assumption that, when faced with competing models, individuals tend to adopt only the model that best fits the observed data for updating their beliefs. This updating rule is prevalent across different environments, including those with and without a specified model prior, when people encounter conflicting or non-conflicting models, and when another model that provides the Bayesian prediction is available. We also uncover other biases in model weighting, particularly one-stage updating, and mistakes in identifying the best-fitting model. Finally, we show that a large share of vectors of posteriors are Bayes-inconsistent, which fundamentally violates Bayesian updating and supports a key prediction from the theoretical literature on model persuasion. These findings not only provide empirical foundations for the assumption in recent papers formalizing narratives as models but also hold important implications for policy-oriented research, which we address in the conclusion.

This paper contributes to the fast-growing literature on narratives in economics, started by Shiller (2017). While there is no consensus in economics on the term narrative (see Barron and Fries, 2024), narratives are mostly formalized in two ways: directed acyclic graphs (DAGs; e.g., Eliaz and Spiegler, 2020, 2024; Eliaz et al., 2021, 2025) or models as introduced by Schwartzstein and Sunderam (2021). Several papers build on the latter framework, and either follow Schwartzstein and Sunderam (2021) by assuming that agents select the best-fitting model (Ichihashi and Meng, 2021; Aina, 2025; Jain, 2024; Schwartzstein and Sunderam, 2024; Bauch and Foerster, 2024), or consider alternative updating rules (Yang, 2023; Ispano, 2025; Wojtowicz, 2024).[2] Building on these assump-

[2]The maximum likelihood criterion is also employed to study a wide range of settings, such as how individuals combine multiple forecasts (Levy and Razin, 2021), or how they face frictions in evaluating the model's fit to data due to computational and cognitive constraints (Samuelson and Steiner, 2024), or how to shift paradigm upon observing an unlikely news (Ortoleva, 2012). In the same spirit, Izzo et al. (2023) formalize models as linear relations between policies and their outcome, assuming that the model with the smallest mean squared error is adopted. By contrast, Barberis et al. (1998) posit that individuals hold competing mental models of asset markets and update beliefs about these models based on stock market signals, following Bayesian updating. This aligns with the literature of psychology which increasingly models cognition as probabilistic inference, often through a Bayesian framework (Tenenbaum et al., 2011; Chater and Oaksford, 2008, for overviews), and justified through the idea of sampling even within individuals (Vul and Pashler, 2008; Vul et al., 2014). Instead, other studies suggest that individuals tend to select the hypothesis with the highest explanatory power (e.g., Douven and Schupbach, 2015a,b).

tions, these papers then study how a persuader, whose preferences might be misaligned with those of a receiver, can use models to manipulate the receiver's beliefs (model persuasion). Our study builds on this framework yet eliminates any strategic considerations. This approach enables us to isolate biases in model weighting from strategic factors, including skepticism and mistrust towards models provided by persuaders.

The empirical literature on narratives has thrived in the last years (e.g., Andre et al., 2022, 2023; Graeber et al., 2024a,b; Bursztyn et al., 2023), with some recent experimental studies exploring related questions tied to the formalized notions of narratives. Barron and Fries (2025) demonstrate in a financial setting that persuaders take model fit into account when proposing a model, and that the fit of the model is important for its persuasiveness. In contrast, we focus on how individuals update their beliefs when confronted with multiple models, distinguishing between different updating rules. Two experimental papers building on DAGs answer questions that relate to ours: Ambuehl and Thysen (2024) analyze the determinants of the selection of causal structures, while Kendall and Charles (2025) examine the effectiveness of causal structures in influencing decisions, uncovering averaging behaviors in the presence of multiple DAGs. In these experiments, participants are provided with rich datasets that maintain fixed correlations between variables, along with one or more DAGs that describe the potential causal relationships among these variables and imply different optimal actions. In contrast, our focus is on belief updating, which depends on the correlations between variables rather than on their causal links. Thus, we present participants with different models that propose different correlations between the variables, keeping the causal structure fixed.[3] Finally, two recent studies using alternative frameworks to analyze model uncertainty provide complementary evidence to our findings: Musolff and Zimmermann (2025) show that computational complexity increases the tendency to rely on model selection, while Augenblick et al. (2025a) find that neglecting uncertainty across multiple models leads to systematic overprecision.

Our paper lies at the intersection of research on narratives and belief updating, thereby also making contributions to the latter. A large and long-standing literature documents biases in belief updating when individuals encounter a single model (for reviews, see Benjamin, 2019; Camerer, 1995). Recent papers have highlighted how features of the information structures (Ba et al., 2025; Augenblick et al., 2025b) and the environment (Bordalo et al., 2026, 2025) might impact the type of biases observed in belief updating, such as over- and underinference or base-rate neglect. The paper most closely related to ours is Liang (2025), which compares belief updating across two settings: individuals are presented with either two data-generating processes that differ in accuracy (uncertain news) or an equivalent data-generating process (certain news) leading to the same Bayesian prediction. His findings indicate that beliefs about the state are less respon-

[3]Relatedly, Fréchette et al. (2024) and Kendall and Oprea (2024) study how people form and extract mental models for these large datasets in the absence of provided DAGs. Fréchette et al. (2024) finds that the recurring mistakes are associated with ignoring or misunderstanding correlations in the data.

sive to the signal in the first setting compared to the second.[4] Rather than focusing on the comparison between certain and uncertain news, our contribution is to study biases in model weighting and their implications for belief updating. We do so by designing an experiment where we can identify the model weights and thus study biases in model weighting by classifying participants' behavior into different updating rules.[5]

The rest of the paper is organized as follows: Section 2 presents the conceptual framework, Section 3 and Section 4 describe the experimental design and results of the main study, Section 5 presents the design and results of the additional studies, and Section 6 discusses the results and concludes.

2 Framework

This section outlines the theoretical framework for understanding how individuals update their beliefs when confronted with conflicting models. Building closely on the theoretical framework of model persuasion, we outline the main updating rules under examination.

Consider a decision-maker (DM, hereafter) with the goal of learning about an unknown binary state of the world, $\omega \in \{A, B\}$, where both states are initially equally likely. The DM observes a signal, $s \in \{p, o\}$, and revises their beliefs about the state. There are two competing models, Model 1 (m_1) and Model 2 (m_2), that could explain the signal and provide predictions about the state. More precisely, each model specifies the probability of each signal conditional on each state, $\Pr^m(s|A)$ and $\Pr^m(s|B)$. By applying Bayes' rule, a model makes predictions about the state for each signal, $\Pr^m(A|s)$.[6]

The DM can then use the observed signal to infer (1) how likely the state is given each model and (2) how likely the models are. The resulting posterior attached to state A conditional on observing signal s can be expressed as a weighted average of the two model predictions over state A:

$$\hat{\Pr}(A|s) = \rho_s^{m_1} \Pr^{m_1}(A|s) + \rho_s^{m_2} \Pr^{m_2}(A|s), \tag{1}$$

with $\rho_s^{m_1}, \rho_s^{m_2} \in [0, 1]$ and $\rho_s^{m_1} + \rho_s^{m_2} = 1$. The weight ρ_s^m corresponds to the weight assigned to model m given signal s.

A large and long-standing literature has focused on how people update beliefs about states

[4] A related pattern of underreaction is documented by Epstein and Halevy (2024) in settings with ambiguous news, where it is interpreted as sensitivity to signal ambiguity.

[5] The relationship between biases in model weighting—the focus of our paper—and biases in inference about the state of the world—the focus of Liang (2025)— is not straightforward, as discussed in more detail in Section 4.7. For example, underinference about the state can be consistent with both underinference about models as well as overinference about models.

[6] This framework can capture the examples discussed in the introduction. In the example about asset markets, the future performance of an asset could either be high ($\omega = A$) or low ($\omega = B$). The signal realizations include information about past performance (e.g., asset prices recently decreased: $s = o$). A model predicting mean reversion would imply $\Pr^m(A|o) > \Pr^m(A)$, whereas a model predicting momentum would imply $\Pr^m(A|o) < \Pr^m(A)$.

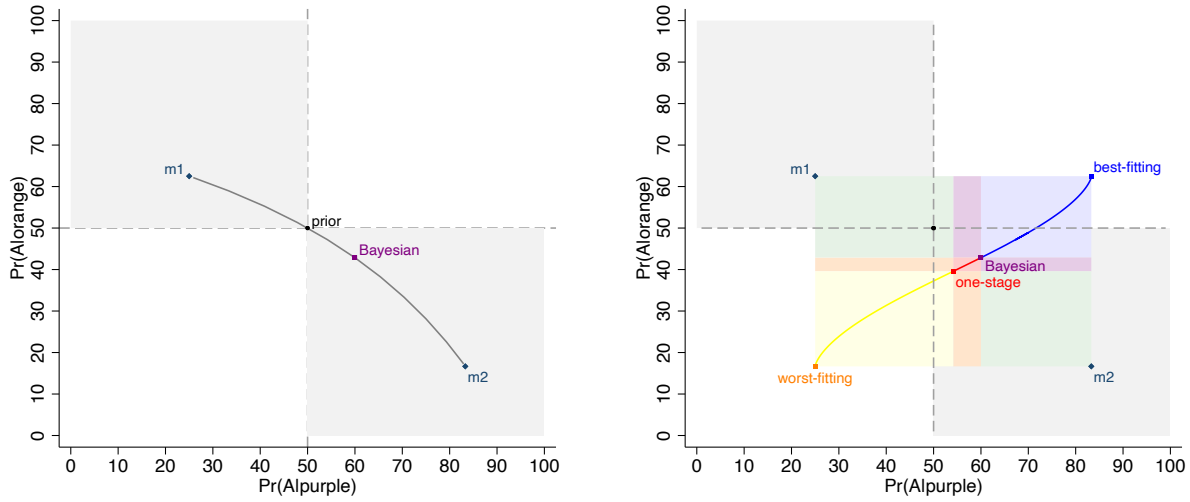
when exposed to one model, which relates to the matter of correctly forming the model predictions, $\Pr^m(A|s)$. In contrast, this paper focuses on how people learn about the models and, in turn, weight the model predictions, ρ_s^m, when making inference about the state. To study model weighting, we assume that the DM correctly derives the Bayesian model predictions. One interpretation of this assumption is that the DM encounters the models together with their predictions, which we view as a realistic feature of relevant settings. We explain in Section 3 how we ensure that this assumption is met in our experimental design.

A Bayesian DM derives the model weights by using Bayes' rule: $\rho_s^m = \Pr(m|s) = \Pr(m)\Pr(s|m)/\Pr(s)$, where $\Pr(m)$ is the prior over model m and $\Pr(s|m)$ is the *likelihood* of observing signal s given model m, which we also refer to as the *fit* of model m given signal s. Such Bayesian model weights have been assumed, for example, by Barberis et al. (1998) to aggregate competing mental models of price movements in asset markets. Figure 1a illustrates the posteriors of a Bayesian DM. In this and similar figures, the axes represent posterior beliefs for the two signal realizations: $\hat{\Pr}(A|p)$ on the x-axis and $\hat{\Pr}(A|o)$ on the y-axis. We refer to points in this graph as *vectors of posteriors*. Each model induces a vector of posteriors corresponding to the model predictions, represented by a blue diamond shape. The DM's prior over the state is represented by the black circle labeled "prior." The vector of posteriors for a Bayesian DM lies on the black line depending on their prior over models. We restrict our attention to the case where the models are initially equally likely, $\Pr(m_1) = 50\%$, represented by a purple square labeled "Bayesian."

While Bayesian updating typically requires the DM to assign positive weight to both models, recent theoretical papers propose an alternative approach—which we call *model selection*—where individuals place full weight on one of the two models, i.e., $\rho_s^m \in \{0, 1\}$. The most common assumption in the literature on model persuasion, as introduced by Schwartzstein and Sunderam (2021), proposes the DM selects the model that maximizes the likelihood of the data, which we refer to as *model selection via maximum likelihood*. Formally, $\Pr(s|m) > \Pr(s|m')$ implies $\rho_s^m = 1$ and $\rho_s^{m'} = 0$, and thus the DM's posterior matches the prediction of the best-fitting model, $\hat{\Pr}(A|s) = \Pr^m(A|s)$.[7] If the process of model selection exhibits some stochasticity, as suggested by Wojtowicz (2024), the DM may occasionally make mistakes in evaluating the fit levels of the competing models and select the worst-fitting model instead of the best-fitting one. Such mistakes differ from the standard way of modeling errors in reported guesses, which posits that DMs make frequent but mostly minor mistakes. In contrast, should errors in model selection occur, they would be of much larger magnitude.

Our main focus is on model selection via maximum likelihood because of its prevalence in theoretical work, but we also consider other criteria for model selection. First, a DM

[7]This approach is equivalent to selecting models according to the Bayes factor, i.e., the ratio of the fit levels of the two competing models $\Pr(s|m)/\Pr(s|m')$. It quantifies the support for one model over the other: if higher than 1, the data supports model m more strongly than m'.



(a) Bayesian Updating

(b) Biases in Model Weighting

Figure 1: Predictions for Vectors of Posteriors

Notes. The figures illustrate two models with the following parameters: $\Pr^{m_1}(p|A) = 1/6$, $\Pr^{m_1}(p|B) = 3/6$, $\Pr^{m_2}(p|A) = 5/6$, and $\Pr^{m_2}(p|B) = 1/6$. In both figures, the black circle corresponds to the prior over the state, the blue diamonds to the two model predictions, the purple square to Bayesian updating with $\Pr(m_1) = 50\%$, the red square to the one-stage updating, the blue square to selecting the best-fitting model, and the orange square to selecting the worst-fitting model; the gray areas represent the Bayes-consistent vectors of posteriors. The black line in Figure 1a illustrates the Bayesian vectors of posteriors for any prior over the models. In Figure 1b, colored areas represent the classification for over- and underinference about the models with the following color code: overinference for both signals (blue), underinference for both signals (red), overinference for a signal and underinference for the other signal (purple), wrong direction for both signals (yellow), overinference for a signal and wrong direction for the other signal (green), and underinference for a signal and wrong direction for the other signal (orange). The colored lines capture overinference (blue), underinference (red), and wrong direction (yellow) for a constant degree of bias in inference about the models, α, as defined by $\rho_s^m = \hat{\Pr}(m|s) = 1/(1 + (\Pr(s|m')/\Pr(s|m))^{\alpha})$ (see Appendix D.1 for details).

may consistently select the same model irrespective of the observed signal, as commonly assumed in the literature on learning with misspecified models (see the discussion in Ba, 2024). This would be the case if the DM applies a model selection criterion based on features of the models that are independent of the signal, such as symmetry or simplicity. Borrowing the term from Ba (2024), we refer to such rules as *dogmatic model selection*. A further alternative is *model selection based on informativeness*. Inspired by Yang (2023), this rule prescribes that the DM selects the most informative model given the observed signal, that is, the model with the posterior over the states closest to 100% or 0%.[8]

Biases in model weighting may not be limited to model selection; a DM that assigns positive weights to both models may do so in a systematically biased manner. In particular, the DM may exhibit over- or underinference about the models. Note that a

[8]Yang (2023) proposes model selection based on decisiveness: individuals choose models that provide clear guidance for optimal action, thus favoring extreme models. This criterion relies on the DM's incentives; thus, decisiveness and informativeness can differ for some payoff structures. Our experiment is not designed to study model selection based on decisiveness, as participants do not make subsequent decisions, and incentives are chosen only to incentivize truthful belief elicitation.

Bayesian DM also revises their beliefs about the best-fitting model m upwards, i.e., $\Pr(m|s) > \Pr(m) = 50\%$. Therefore, any $\rho_s^m \in (\Pr(m|s), 1]$ would reveal overinference about the models: the DM overreacts to the observed signal, expecting model m to be the true data-generating process with an excessively high probability. With $\rho_s^m = 1$, model selection via maximum likelihood corresponds to the extreme form of overinference about the models. Conversely, a DM may exhibit underinference about the models if $\rho_s^m \in [0.5, \Pr(m|s))$. In its extreme form, the DM weights the two model predictions using the prior over the models, $\rho_s^m = \Pr(m) = 50\%$. We refer to this as *one-stage updating* because the DM recognizes that the signal is informative for learning about the state given each model but fails to update beliefs about the models, improperly combining the model predictions. Such an updating rule is assumed by Mullainathan et al. (2008) to study persuasion with categorical thinking. To capture more broadly over- and underinference about the models, we present in Appendix D.1 a theoretical framework that can account for less extreme forms of over- and underinference about the models.

Figure 1b illustrates the predictions for the different biases in model weighting. The squares labeled "Bayesian" and "one-stage" correspond to the predictions of Bayesian and one-stage updating. A DM exhibiting overinference (underinference) about the models for both signals would report a vector of posteriors in the blue (red) area. The blue square labeled "best-fitting" corresponds to the prediction of model selection via maximum likelihood. If the DM makes mistakes in applying this rule, their reported vectors of posteriors may coincide with the predictions of Model 1 ("m1"), Model 2 ("m2") or the worst-fitting model ("worst-fitting"). A vector of posteriors at "m1" or "m2" might also indicate dogmatic model selection, while model selection based on informativeness would imply reporting "m2."

Biases in model weighting can lead to patterns that are highly inconsistent with Bayesian updating. For instance, the DM might become more confident in one state regardless of the observed signal, as illustrated by the "best-fitting" prediction in Figure 1b. We say the reported vector of posteriors is *Bayes-consistent* if, for all $s \neq s'$, (i) $\hat{\Pr}(A|s) > \Pr(A)$ and $\hat{\Pr}(A|s') < \Pr(A)$, or (ii) $\hat{\Pr}(A|s) = \hat{\Pr}(A|s') = \Pr(A)$. Bayes consistency is the sole restriction that Bayesian updating generally imposes on vectors of posteriors, irrespective of the set of plausible models the DM may entertain (Aina, 2025).[9] However, for model selection via maximum likelihood, each signal triggers the adoption of a different model, and, as a result, *Bayes-inconsistencies* occur every time the DM is confronted with conflicting models, that is, pairs of models for which $\Pr^m(s|A) > \Pr^m(s|B)$ and $\Pr^{m'}(s|A) < \Pr^{m'}(s|B)$. In Figure 1, the gray areas correspond to Bayes-consistent vectors of posteriors, while the white areas highlight the Bayes-inconsistent ones. The two illustrated models are conflicting as their predictions fall on opposing gray areas and,

[9]Following Aina (2025), Bayes-consistency requires that the prior can be expressed as a convex combination of the posteriors across signals for each state; see Shmaya and Yariv (2016) and Bohren and Hauser (2024) for analogous results. Note that Bayes-consistency is a weaker condition than requiring a distribution of posteriors to be Bayes-plausible as in Kamenica and Gentzkow (2011). In Figure 1, the only vector of posteriors consistent with Bayes-plausibility is the Bayesian prediction.

indeed, the best-fitting prediction lies in the white area.

Bayes-inconsistencies do not only result from model selection via maximum likelihood. Rather, they occur frequently when DMs are faced with competing models and exhibit strong overinference about the models. In contrast, when a DM entertains only a single model, there is no reason to expect Bayes-inconsistencies to arise. This insight is also important for distinguishing model weighting from another possible updating procedure: instead of weighting the two model predictions, the DM could aggregate the models into a single model and then update their beliefs using this model. According to this approach, which we discuss in detail in Appendix D.3, the resulting posterior beliefs would be Bayes-consistent—even if the DM's belief revision exhibits biases in updating about the state—since they are derived using a single model. Therefore, Bayes-inconsistencies offer compelling evidence of novel biases in belief updating that arise when confronted with multiple models. Moreover, these inconsistencies highlight that biases in model weighting can lead to deviations from the Bayesian benchmark that fundamentally differ from those observed when updating with a single model.

This section highlights that the literature has considered a wide range of plausible model weighting rules. Distinguishing between these rules has important implications. For example, Bayesian inconsistencies only arise from rules that result in strong overinference about the models. This underscores the need for empirical evidence to determine how individuals actually update in these situations, which is the purpose of our study. We examine the descriptive validity of Bayesian updating, one-stage updating, and model selection using different selection criteria. Specifically, we investigate which rules are applied frequently and whether individuals apply them consistently across updating tasks. Finally, we also study whether Bayesian inconsistencies occur frequently.

3 Experimental Design

The focus of our experiment is to examine how people update beliefs in the presence of competing models. To address this question and comprehensively study biases in model weighting, our experimental design meets the following criteria: (i) using a belief updating task where participants can be exposed to multiple models that could have generated the data and that closely follows the theoretical framework; (ii) controlling the true data-generating process and the pairs of models participants encounter; (iii) avoiding biases and errors in deriving models' predictions; and (iv) collecting the vectors of posteriors.

To achieve features (i) and (ii), we build on the classic "balls-and-urns" paradigm (Edwards, 1968). Employing this context-free setting also allows us to rule out confounds related to pre-existing beliefs or experiences. Regarding feature (iii), we provide participants with the model predictions, as detailed below. This approach eliminates biases in deriving model predictions, enabling us to directly measure biases in model weighting. Without this feature, identifying biases in model weighting would be challenging. This

relates to Bohren and Hauser (2024), which highlights the challenges of distinguishing biases from non-Bayesian updating and those from the use of misspecified models. An alternative approach would be to directly elicit model weights rather than inferring them from posteriors over the state, but this would fundamentally alter the updating environment that motivated our study.[10] To implement feature (iv), participants encounter some updating tasks twice, allowing us to recover the vectors of posteriors. This is important for gaining deeper insights into the updating process and for better distinguishing among potential biases in model weighting. Additionally, it allows us to test for the systematic presence of Bayes-inconsistencies, a key consequence of certain biases in model weighting.

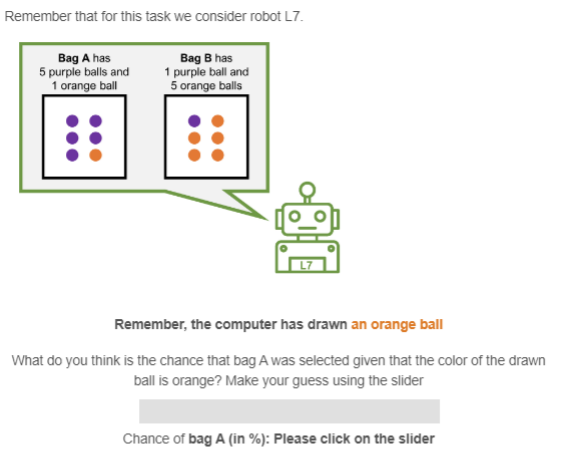
In the classic balls-and-urns paradigm, the participant encounters one model. There are two bags, which represent the state of the world, $\omega \in \{A, B\}$, and each contains a number of colored balls. In our experiment, bags always have a total of six balls, with each ball being either purple, p, or orange, o. A model m sets the number of purple balls in each bag, $\Pr^m(p|A)$ and $\Pr^m(p|B)$. The participants know the composition of both bags given the model. In the task, one of these bags is randomly selected. The participants do not know which bag is selected but observe the color of a ball drawn from the selected bag, that is, the signal $s \in \{p, o\}$. After observing the signal, the participants report their beliefs about the probability that bag A was selected.

To study model weighting, we introduce a novel updating task that closely follows the theoretical framework, building on the classic balls-and-urns paradigm. Before detailing the task, we outline the key idea. In this novel task, participants face two models. Participants learn the compositions of the bags according to both possible models before they observe the signal, as in Aina (2025). In this updating task, one of the two models is randomly selected to be the true data-generating process and to generate the observed signal. That is, the selected model dictates the composition of the balls in the bags. The remaining sequence follows the standard updating task: a bag is selected, and then a ball is drawn from that bag. The participants know neither the selected model nor the selected bag. After observing the color of the drawn ball, the participants are asked to report their beliefs about bag A being selected.

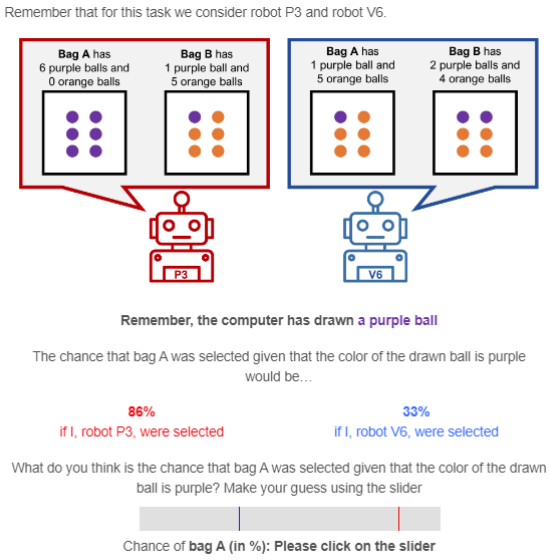
To introduce this novel setting to participants, we divide our experiment into three parts, progressively incorporating new elements through a series of updating tasks. We first introduce updating with a single model (Part 1), then the model predictions (Part 2), and, finally, updating with two models (Part 3). This helps participants to better understand the different elements of our novel updating task. Three details regarding the implementation are worth mentioning. First, we use an animated interface illustrating the random

[10]We choose to elicit posteriors over the state because we want to closely mimic both the theoretical framework and the natural decision-making process, in which individuals use models as means to update their beliefs about the state in response to new information. Directly eliciting model weights would shift the focus to a setting where individuals are interested in learning about models themselves, thereby eliminating the central role of the payoff-relevant state of the world—a substantially different decision-making context. Moreover, we believe that directly eliciting model weights may communicate to participants that we expect them to react to the signal in a particular way.

draw of the state, the draw of the signal, and, in Part 3, the draw of the model, to recreate a realistic setting online and intuitively remind the participants of the task's basic structure. Second, we framed models as "robots" placing a specific number of orange and purple balls in each bag. The goal is to make the task more intuitive, especially in Part 3, where the participants encounter two models. Third, participants report their guesses using a slider. Next, we provide a more detailed discussion of each of the three parts.



(a) Part 1



(b) Part 3

Figure 2: Experimental Interface

Part 1 employs classic balls-and-urns updating tasks with one model. The purpose of this first part is to introduce and explain updating in the context of a single model. Participants make guesses for two different models, described in Table 1. Figure 2a illustrates the survey interface.

Part 2 mirrors the first part with one variation: the robot provides the participant with the correct Bayesian predictions, $\Pr^m(A|s)$, given the observed ball. This part introduces and explains model predictions. Predictions are conveyed both verbally and through a colorful pin on the slider. Participants then make two guesses with the same models as in Part 1. Since the correct Bayesian predictions are provided, these updating tasks only serve as a learning exercise.

Part 3, which is the focus of our study, alters the paradigm by introducing two robots, each proposing a different distribution of colored balls in the bags. With equal probability, one of the two robots is chosen to implement its proposed distribution. Hence, participants are presented with two models. Similar to Part 2, each robot provides the Bayesian prediction given the observed drawn ball. Figure 2b illustrates the survey interface (see also Appendix G). Note that this interface allows for a simple frequentist way of computing the Bayesian prediction under model uncertainty (Gigerenzer and Hoffrage,

1995).[11]

In Part 3, participants complete seven updating tasks. We carefully chose a total of five pairs of models, presented in Table 1. When choosing the model pairs, we focused on those that (i) have different fit levels, ensuring a single best-fitting model, (ii) are conflicting, so that model selection via maximum likelihood predicts Bayes-inconsistent vectors of posteriors, and (iii) allow us to distinguish between the different updating rules discussed in Section 2, (iv) exhibit diverse characteristics to assess the robustness of our findings. Appendix Figure B1 shows the predictions of the different updating rules for all model pairs. Moreover, Model Pair 3 and Model Pair 5 share a common model, which allows us to test, on aggregate, whether reported posteriors are reactive to making the other model's prediction more informative.[12]

Participants encounter these five model pairs in a random order, followed by a repetition of two model pairs (Model Pair 2 and Model Pair 3), also presented in a random order (see Table 1, Column "Repeated"). We use the data from the repeated model pairs in two ways. First, when participants report posteriors for the same model pair observing different signals, we observe their resulting vector of posteriors, $\hat{\Pr}(A|p)$ and $\hat{\Pr}(A|o)$.[13] Second, when participants report posteriors for the same model pair observing the same signals, we can use these observations to assess data quality by examining the consistency of reported guesses and to study stochasticity in applying model selection criteria.

We conclude the study by eliciting several survey items, including three contextualized updating tasks, a Berlin numeracy task (Cokely et al., 2012), a modified version of the Cognitive Reflection Test (Frederick, 2005), a selection of items to determine their thinking style (Keaton, 2017), and demographics.

Incentives The participants earned a completion fee of 6 USD and, depending on the accuracy of their guesses in the updating tasks, could earn a bonus of 2 USD. Only one of the eleven updating tasks was randomly selected for the bonus payment. Belief elicitation is incentivized using the binarized scoring rule (Hossain and Okui, 2013) and explained intuitively by following Danz et al. (2022).[14]

[11] With the uniform prior over the state and models, participants need only compare the number of balls of the drawn color in Bag A (summed across both robots) to the total number of balls of that color in both bags (summed across both robots). For example, in Figure 2b, a purple ball is drawn. There are 7 purple balls in Bag A across the two models (6 from Robot P3 and 1 from Robot V6), and 10 purple balls in total (7 from Robot P3 and 3 from Robot V6). Thus, the Bayesian prediction is $7/10 = 70\%$.

[12] Model Pair 4 is designed to compare two particular models: one fully uninformative (same number of purple balls in each bag) and one fully informative (only one color of balls in each bag). This pair offers a particularly strong test of model selection based on informativeness, which would predict a systematically higher weight towards the fully informative model regardless of the observed signal.

[13] We opt for repeated model pairs rather than eliciting vectors of posteriors directly. To do that, one would need to elicit beliefs with the strategy method; however, Aina et al. (2025) finds reported beliefs might differ if reported contingent on both signal realizations, without having observed the realized one.

[14] Instructions clarify that "to maximize the chance of winning the bonus, it is in your best interest always to give a guess that you think is the true chance." A control question verifies the comprehension of this aspect. Participants have the option to review the details of the elicitation rule at their discretion.

Part	Pair	$\Pr^{m_1}(p\|A)$	$\Pr^{m_1}(p\|B)$	$\Pr^{m_2}(p\|A)$	$\Pr^{m_2}(p\|B)$	Repeated
1, 2		5/6	1/6			
1, 2		4/6	5/6			
3	1	1/6	2/6	5/6	2/6	No
3	2	1/6	3/6	5/6	1/6	Yes
3	3	6/6	1/6	1/6	2/6	Yes
3	4	1/6	1/6	0/6	6/6	No
3	5	4/6	2/6	1/6	2/6	No

Table 1: Models Used in the Experiment

Notes. This table describes all models used in the experiment. In Part 1 and Part 2, participants encounter a single model (m_1), while in Part 3, they face two models (m_1 and m_2). In each updating task, there are two bags, representing the state of the world, $\omega \in \{A, B\}$, with each bag containing six balls, either purple (p) or orange (o). A model m sets the share of purple balls in each bag, $\Pr^m(p|A)$ and $\Pr^m(p|B)$, as shown in the corresponding columns. Finally, in Part 3, participants encounter Model Pair 2 and Model Pair 3 twice, as indicated in the "Repeated" column.

Logistics and Sample The experiment was pre-registered on AsPredicted and conducted on Prolific in April 2024, restricting the participant pool to US residents, aged 18-70 with approval rates of at least 95%.[15] The study was completed through a link to a Qualtrics survey, including instructions and control questions for each part (see Appendix G). A total of 300 participants completed the study successfully. The average payment was 7.5 USD, and the average duration was approximately 38 minutes. In our final sample, 51% are female, 31% have low schooling ("High school" or lower educational level), and the median age is 38.

4 Results

In this section, we analyze how individuals update beliefs when confronted with competing models. We analyze the data in multiple forms. First, we pool the data from all seven updating tasks and study individual guesses. Second, we investigate the reported vectors of posteriors to better distinguish between the predictions of different update rules and to test for the systematic presence of Bayes-inconsistencies. Third, we study whether participants consistently apply the same updating rules across all seven updating tasks. We begin by focusing on the main updating rules introduced in Section 2 and present results for additional potential updating rules in a subsequent section. Before presenting these analyses, we begin by evaluating the quality of the collected data.

4.1 Quality of Data

We consider two complementary methods to verify the quality of our data: assessing individual consistency when participants repeatedly face the same updating task and analyzing the aggregate reaction across model pairs.

[15]The pre-registration plan is available at `https://aspredicted.org/9YD_HNH`.

Consistency First, we examine the consistency of guesses when participants face the same updating task. We focus on the repeated model pairs and analyze instances where participants observed the same signal twice.[16] This approach enables us to assess the noise in the reported guesses in our experiment. Our data reveal that a substantial portion of the guesses are consistent; specifically, 42% of these pairs of guesses are identical, and an additional 11% differ by only a small margin of at most 2 percentage points. The average distance between guesses is 11%, with a median of 1%. Appendix Figure A1 reports the cumulative distribution of this measure.

Across Model Pairs Second, we compare the reported guesses across two model pairs, Model Pair 3 and Model Pair 5, which share a common model. This approach allows us to test whether, on aggregate, participants are responsive to the information provided in our experiment by comparing their average beliefs across model pairs when making a model more informative for both signal realizations. If this were the case, we would expect that a DM reports a higher (lower) average guess when a purple (orange) ball is drawn in Model Pair 3 than in Model Pair 5 (see Appendix Figure A2).[17] Both these predictions are verified in our data. The average guesses given a purple ball are 67% for Model Pair 3 and 53% for Model Pair 5 (p-value < 0.001), and average guesses given an orange ball are 44% for Model Pair 3 and 50% for Model Pair 5 (p-value $= 0.001$).

The high consistency rate and aggregate reactions to changing a model across model pairs indicate high data quality and suggest that participants understand the choices they face.

4.2 Analysis of Individual Guesses

In this section, we analyze the individual reported guesses, pooling the data across all seven updating tasks. Figure 3 presents the distribution of the weights assigned to the best-fitting model, ρ, based on participants' guesses. The blue distribution shows the recovered weights for the reported guesses that fall within the range of the two model predictions (87.86% of all guesses, N=1,845), while the red distribution serves as a benchmark, illustrating how the weights would be distributed if participants adhered to Bayesian updating in all tasks. The figure shows that the distribution of ρ deviates significantly from the Bayesian benchmark, with prominent peaks at 100%, 0%, and 50%, corresponding to the predictions of selecting the best-fitting model or the worst-fitting model, and one-stage updating, respectively. The absence of substantial peaks at other weights suggests that participants seldom employ alternative updating rules that would require a constant

[16] We can perform this test for a total of 300 pairs of guesses, corresponding to 220 participants: for 80 participants, different signals were drawn for both repeated model pairs, and we cannot perform this test; for 140 participants, we have the pair of guesses for one model; and for 80 participants, we have the pair of guesses for both model pairs.

[17] Such updating behavior is predicted by most of the updating rules that we consider, including Bayesian updating, one-stage updating, and model selection via maximum likelihood. Only dogmatic model selection does not make specific predictions in this setting, as it imposes no structure on model selection beyond the requirement that participants choose the same model for both signals.

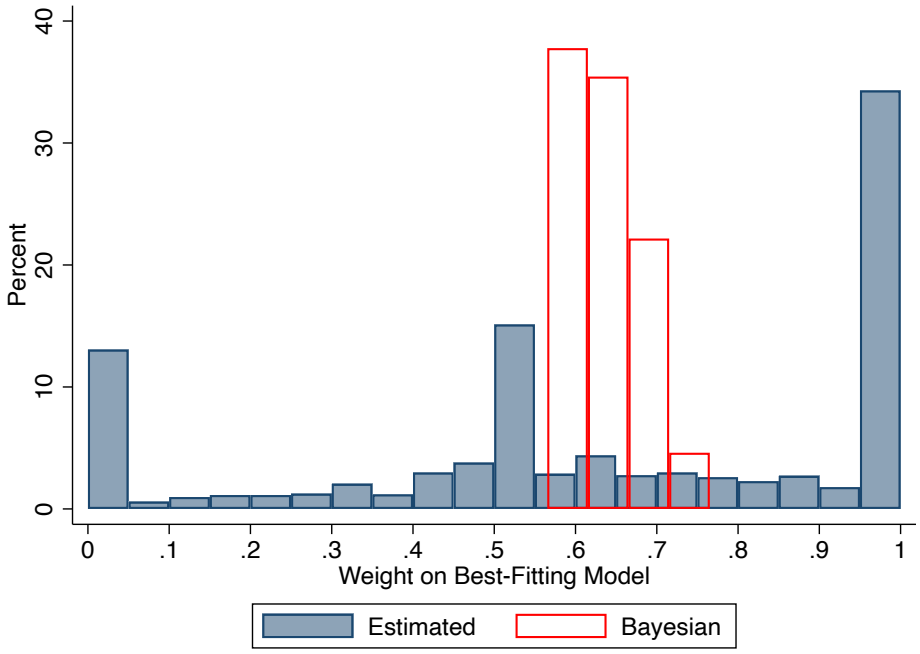


Figure 3: Estimated Model Weights

Notes. The figure plots the distribution of weights on the best-fitting model, ρ, across tasks. We report the implied weights for the reported guesses within the two model predictions in blue. We compute the implied weight as $\rho = (guess - \Pr^{m'}(A|s))/(\Pr^{m}(A|s) - \Pr^{m'}(A|s))$, where m represents the best-fitting model and m' the worst-fitting model given signal s. The red distribution serves as a benchmark, illustrating how the weights should be distributed if participants follow Bayesian updating in all tasks.

ρ across tasks.[18]

Interestingly, the mean reported guess is relatively close to the Bayesian benchmark, with an average deviation of 4.98 percentage points. While this difference is small, it is statistically significant (OLS regression with cluster-robust standard errors: $t = 8.55$, $p < 0.001$). Appendix Figure A3 plots the average reported guess alongside the Bayesian prediction for each of the five model pairs and two signal realizations, revealing similar patterns. However, the belief distributions are multimodal, with few observations near the Bayesian prediction (see Figure 3 and Appendix Figures B2, B3 and B4). This highlights that while mean beliefs seem to align with Bayesian predictions, they fail to adequately capture the population's belief distributions—a pattern also emphasized in Bordalo et al. (2026).

We further classify the guesses by the point predictions of the different updating rules discussed in Section 2. Table 2 reports the percentage of guesses classified as Bayesian, one-stage, and model selection.[19] For model selection, we differentiate between selecting the best-fitting and the worst-fitting model. Strikingly, by only considering these four

[18]We explore updating rules with constant model weights in Appendix D.2 and find little support for this behavior. We discuss these and other alternative updating rules in Section 4.5.

[19]Appendix Table A1 replicates this classification for consistent guesses, as described in Section 4.1. The overall pattern closely resembles the one of Table 2, but with a higher share of guesses matching the prediction of the best-fitting model (48%) and less falling in the residual category "Within" (20%).

Type of Guess	Exact		Within 2 p.p.	
	%	95%-CI	%	95%-CI
Bayesian	3.05	[1.48, 4.61]	8.14	[6.20, 10.09]
One-stage	10.95	[8.21, 13.69]	18.00	[14.61, 21.39]
Best-fitting	28.86	[25.03, 32.68]	33.10	[29.14, 37.05]
Worst-fitting	10.95	[8.88, 13.03]	12.43	[10.26, 14.60]
Within	34.05	[30.16, 37.94]		
Outside	12.14	[9.94, 14.35]		
Total	100.00		71.67	

Table 2: Classification of Guesses

Notes. Column "Exact" reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, and worst-fitting) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise). Column "Within 2 p.p." reports the shares of guesses that fall within 2 p.p. around each point prediction. Note that 2.33% of reported guesses are classified both as Bayesian and one-stage, not included in the total. Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We pool the data from all seven updating tasks.

point predictions, we classify over half of all guesses when requiring an exact match to the predictions (Column "Exact") and more than 70% of guesses when allowing for a small deviation of at most 2 percentage points from the predictions (Column "Within 2 p.p."). In Column "Exact", we classify guesses that do not match any point prediction into two residual categories: "Within" if the guess falls within the model predictions and "Outside" otherwise. Reassuringly, only 12.14% of guesses are classified as outside, with only 8.33% deviating substantially from the model predictions (more than 2 percentage points). Among all outside guesses, 77% are extreme in the direction of the best-fitting prediction, while 23% are extreme in the direction of the worst-fitting prediction.[20]

The most frequent updating rule used in our data is model selection via maximum likelihood, accounting for about one-third of all guesses. The second and third largest categories are one-stage updating and selecting the worst-fitting model. Some guesses are consistent with Bayesian updating but these are less common; they account for about 8% of all guesses. The confidence intervals provided in Table 2 indicate that our estimates for the different shares are fairly precise. While Table 2 pools the data from all model pairs, Appendix Table B1 and B2 replicate this analysis for each model pair. As discussed in more detail in Appendix B, we find that the distributions of updating rules recovered from individual guesses are remarkably similar across different model pairs.

Overall, model selection is widespread and it accounts for 45.53% of reported guesses. The findings in Table 2 show that participants predominantly update their beliefs using

[20]By adopting a broader classification of over- and underinference about the models, our analysis reveals that the most frequent bias is overinference, with 43% of guesses falling into this category. Furthermore, we find that 17% of guesses are classified as underinference about the models, while 24% represent inferences in the wrong direction. Note that inferences in the wrong direction include selecting the worst-fitting model. In particular, the residual category "Within" of Table 2 can be further divided into 17% underinference, 42% overinference, and 41% wrong direction.

Dependent Variable	Selects Model 1				
	(1)	(2)	(3)	(4)	(5)
Best-fitting	0.450***		0.449***		0.446***
	(0.041)		(0.043)		(0.043)
Most Informative		-0.151***	-0.005		-0.002
		(0.039)	(0.035)		(0.080)
Model Pair 2				0.148***	0.111***
				(0.048)	(0.041)
Model Pair 3				-0.009	0.049
				(0.060)	(0.091)
Model Pair 4				0.035	-0.011
				(0.066)	(0.064)
Model Pair 5				-0.038	0.021
				(0.065)	(0.077)
Constant	0.257***	0.540***	0.259***	0.446***	0.214***
	(0.027)	(0.023)	(0.034)	(0.045)	(0.041)
Observations	836	836	836	836	836
R^2	0.203	0.021	0.203	0.020	0.211

Table 3: Criteria for Model Selection

Notes. The table shows coefficient estimates from linear regressions of an indicator of selecting Model 1 on different explanatory variables. The sample consists of the 836 guesses that correspond to model selection (see Table 2). "Best-fitting" is an indicator for Model 1 having a higher fit than Model 2 given the observed signal. "Most Informative" is an indicator for Model 1 being more informative about the state than Model 2 given the observed signal. "Model Pair 2" to "Model Pair 5" are model pair fixed effects that capture any factors that are unconditional on the signal. Standard errors are clustered on the individual level (179 clusters) and are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the best-fitting model, providing evidence in support of model selection via maximum likelihood. But do participants also use other criteria to select models? To explore alternative criteria, we consider whether participants tend to select the most informative model. We find that around 18% of guesses align with model selection based on informativeness. Importantly, however, depending on the model pair, the predictions of the different model selection criteria overlap. Guesses that correspond to the most informative model are more frequent when the most informative model aligns with the best-fitting model (30%) than when it aligns with the worst-fitting model (11%).

To account for the overlaps in predictions, we regress an indicator of whether the participant selected Model 1 rather than Model 2 on model characteristics corresponding to different model selection criteria. For this analysis, we restrict our sample to observations corresponding to model selection. The results are presented in Table 3. Column (1) shows that participants are 45 percentage points more likely to select Model 1 when

it is the best-fitting model, consistent with model selection via maximum likelihood. Columns (2) and (3) demonstrate that informativeness does not robustly predict model selection.[21] Finally, we investigate whether participants select models based on characteristics of models or of the model pair that are independent of the signal, thus aligned with dogmatic model selection. Columns (4) and (5) explore this by including model pair fixed effects. Most fixed effects coefficient estimates are small and not statistically significant, except for Model Pair 2, where participants seem to favor Model 1 regardless of the signal. However, the R^2 changes minimally with the addition of these fixed effects, suggesting that the signal-independent characteristics do not play a major role in our setting. These results speak against a systematic form of dogmatic model selection, where most individuals consistently select the same model, but they do not rule out an unsystematic form of dogmatic model selection, where participants are drawn to different models in a balanced manner.

Therefore, our findings provide strong support for the importance of the maximum likelihood criterion, while offering little support for the informativeness and a systematic form of the dogmatic criteria, which exhibit little predictive power. Appendix Table A2 presents similar findings, allowing for a difference of at most 2 percentage points between guesses and predictions.

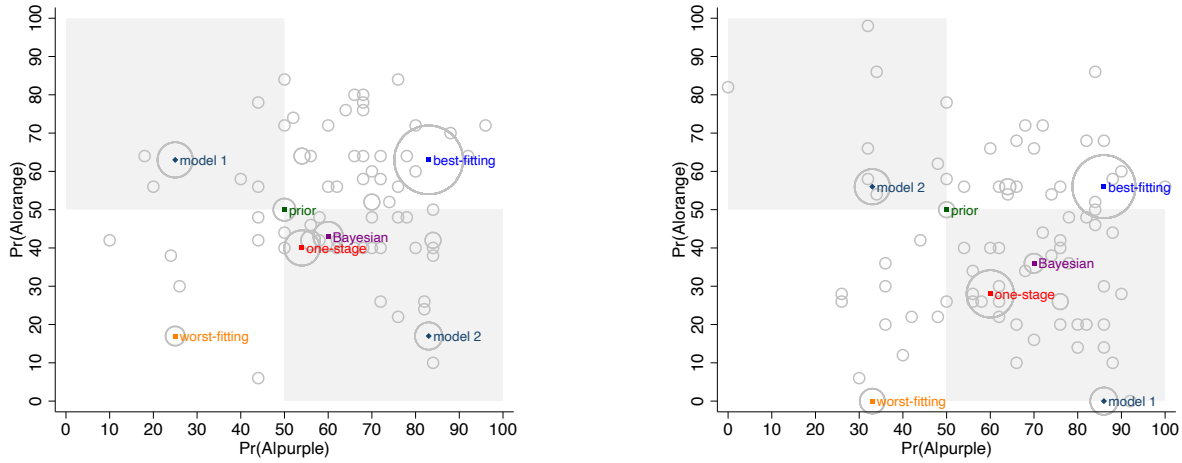
4.3 Analysis of Vectors of Posterior Beliefs

Next, we examine the reported vectors of posterior beliefs. We have data on the vectors of posteriors from the repeated model pairs, when participants observed different signals for the same pair of models. This data enables us to determine whether participants consistently use the same updating rule across different signals, to better distinguish different rules—especially for model selection—and to test for the presence of Bayes-inconsistent vectors of posteriors.

Our analysis is based on a total of 300 vectors of posteriors (148 in Model Pair 2 and 152 in Model Pair 3), gathered from 220 participants.[22] Figure 4 plots the distribution of the reported vectors of posteriors. Table 4 categorizes these vectors of posteriors according to the point predictions discussed in Section 2, pooling together data from both model pairs. Column "Exact" reports the shares of the reported vectors of posteriors that correspond exactly to those predicted by each updating rule. This analysis is considerably more stringent than the one in Section 4.2 as it requires a match with the point prediction for both signal realizations. To relax this demanding requirement, we also look at the share of observations for which the Euclidean distance between the reported vector of

[21] As an additional test of model selection based on informativeness, we can restrict our attention to Model Pair 4, where participants encounter one fully informative (Model 1) and one fully uninformative model (Model 2). We do not observe participants to be systematically drawn to a specific model as shown in Column (4).

[22] For 80 participants, the same signal was drawn for both repeated model pairs, preventing us from recovering any vector of posterior beliefs. Note, however, that this attrition is purely random.



(a) Model Pair 2

(b) Model Pair 3

Figure 4: Vectors of Posteriors

Notes. The size of the circles is relative to the number of observations. We pooled observations that are within 2 p.p. of the point predictions. Observations in the gray areas are Bayes-consistent, while observations in the white areas are Bayes-inconsistent.

Type of Vector	Exact		Within 2 p.p.		Within 5 p.p.	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	2.67	[0.42, 4.91]	3.33	[0.92, 5.75]	5.33	[2.21, 8.46]
One-stage	6.00	[2.88, 9.12]	9.67	[5.77, 13.56]	12.67	[8.24, 17.09]
Best-fitting	19.67	[14.58, 24.76]	23.33	[17.87, 28.80]	24.67	[19.09, 30.24]
Worst-fitting	2.33	[0.61, 4.06]	2.67	[0.83, 4.50]	2.67	[0.83, 4.50]
Model 1	4.33	[1.87, 6.79]	5.33	[2.66, 8.01]	5.67	[2.93, 8.41]
Model 2	4.67	[1.96, 7.37]	6.00	[3.04, 8.96]	6.00	[3.04, 8.96]
Prior	2.00	[0.00, 4.05]	2.00	[0.00, 4.05]	2.00	[0.00, 4.05]
Total	41.67		52.33		59.01	

Table 4: Classification of Vectors of Posteriors

Notes. The table reports the shares of vectors of posteriors that have an Euclidean distance from the prediction vector of posteriors or either 0 (Column "Exact"), 2 p.p. (Column "Within 2 p.p.") or 5 p.p. (Column "Within 5 p.p."). Note that in Column "Within 5 p.p.," 0.33% of reported guesses are classified both as Bayesian and one-stage. We pool the data from both repeated model pairs.

posteriors and the prediction vector is within 2 or 5 percentage points (Column "Within 2 p.p." and Column "Within 5 p.p."). These classifications capture between 42% and 59% of the reported vectors of posteriors.[23]

Despite using substantially more demanding criteria for classification, the data pattern remains closely aligned with the findings we report in our analysis of individual guesses. When allowing for a distance of at most 2 percentage points between the reported vector and the prediction, almost 40% of all reported vectors of posteriors are consistent with model selection for both signal realizations ("Best-fitting," "Worst-fitting," "Model 1," or "Model 2" in Figure 4 and Table 4). Additionally, the two most common predictions continue to be selecting the best-fitting model (23.33%) and one-stage updating (9.67%), while the vectors of posteriors consistent with Bayesian updating are rather rare (3.33%). Also, a few participants consistently report the prior over the state (2%).[24] We do not find evidence of any other frequent and consistent patterns in the data.

Analyzing the reported vectors of posteriors is also valuable in better distinguishing the different criteria for model selection. While selecting the best-fitting model for both signals is common, supporting model selection via maximum likelihood, consistently selecting the worst-fitting model is rare. This suggests that such behavior is more likely a mistake in applying a model selection criterion rather than a systematic updating rule. Dogmatic model selection would imply that the same model is selected for both signals. We only find a few instances in which participants select the same model twice ("Model 1" or "Model 2" in Figure 4 and Table 4), indicating that the dogmatic criterion does not play a major role. Also, the frequencies with which each of the two models is selected are similar, indicating no systematic preference for one of the two models. The informativeness criterion predicts that participants should select Model 2 for both signals in Model Pair 2 and Model 1 for both signals in Model Pair 3. Only 4.05% and 3.95% of vectors of posteriors are consistent with these predictions, respectively, suggesting this is not a predominant model selection criterion.

Selecting the same model for both signals could also be consistent with a stochastic version of model selection via maximum likelihood. We explore this possibility in Appendix C. First, to investigate the stochasticity in model selection, we examine the repeated tasks where participants observe the same signal. Among participants who select a model in both updating tasks (N=118), 16.1% select two different models. These inconsistencies imply that, on average, participants make a mistake in applying their model selection criterion with a probability of 8.83%. Next, we use this estimate and the reported vectors of posteriors to determine the shares of participants following the different model selection criteria but making mistakes with the estimated probability. Among participants that

[23] Appendix Tables A3 and A4 replicate this classification separately for each model pair and show that the distributions are closely aligned. These classifications capture between 41% and 60% of the reported vectors for Model Pair 2, and between 43% and 60% of the reported vectors for Model Pair 3.

[24] We do not include reporting the prior over the state as a category in Table 2 because it overlaps with other predictions for some model pairs. Overall, 6% of the reported guesses correspond to the prior, but only 2.8% for model pairs where there is no overlap with other updating rules.

select models across signal realizations, we estimate that 73.10% follow the maximum likelihood criterion, 21.02% use either the dogmatic or informativeness criteria, and the residual 5.88% aim to consistently select the worst-fitting model. Therefore, this simple calibration exercise indicates that model selection in our data can be well described by a stochastic version of model selection via maximum likelihood, with a small minority of participants using other selection rules.

Importantly, analyzing the reported vectors of posteriors also allows us to test for the presence of frequent Bayesian inconsistencies, meaning that posteriors reported across both signal realizations are consistently higher or lower than the prior. Such inconsistencies are a rather extreme deviation from Bayesian updating and are a crucial consequence of biases in model weighting that systematically assign disproportionate weight to the best-fitting or worst-fitting model. Figure 4 shows that a substantial share of the reported vectors of posteriors are Bayes-inconsistent, as they lie in the gray areas. When pooling data from both model pairs, we find that 50% of vectors are Bayes-inconsistent, of which 41% are in the direction of the best-fitting model and 9% in the direction of the worst-fitting model.[25] Hence, we find compelling evidence for the presence of Bayes-inconsistencies, mostly driven by maximum likelihood selection. Section 5.3 presents further evidence from an additional study that the large share of inconsistent posteriors across signals are driven by systematic biases in model weighting rather than noise.

4.4 Consistency in the Use of Updating Rules

In the previous sections, we observed substantial heterogeneity in the reported guesses; however, a few rules can describe the majority of the reported posteriors: selecting one model, as well as the Bayesian and one-stage updating. In this section, we investigate whether participants consistently use any of these rules across tasks.

We have already presented some evidence of consistent application of updating rules in our analysis of vectors of posterior beliefs.[26] In this section, we extend our analysis to consider all seven updating tasks each participant completes. Table 5 shows how often participants use specific rules, allowing for a distance of at most 2 percentage points between the prediction and the reported guess. We also consider the possibility of reporting the prior over the state, as we see some evidence for such guesses in Figure 4. The table indicates that a large share of participants consistently adhere to one of these updating rules. When requiring participants to use the same rule for all seven guesses, we classify 40.34% of our sample. By allowing participants to deviate from the rule in one or two guesses, we classify 49% and 54% of participants, respectively. We can classify even

[25] In Model Pair 2 and in Model Pair 3, respectively, 53% and 46% of the reported vectors are Bayes-inconsistent. This difference is not statistically different (p-value = 0.206).

[26] Relatedly, we can also investigate the consistency of the guesses from the repeated tasks where the same signal is observed, as in Section 4.1. Appendix Table A5 reports this analysis and documents substantial consistency. Moreover, these estimates are in line with both the shares of guesses and vectors of posteriors reported in the previous two sections. For example, among these pairs of guesses, 28.7% corresponds to selecting the best-fitting model and 11.7% to one-stage updating.

Nr. Consistent Observations	Bayesian	One-stage	Model Selection	Prior
0	67.67	56.33	33.00	49.00
1	21.00	18.33	13.00	39.00
2	7.00	8.67	6.00	8.67
3	1.67	4.67	4.00	1.67
4	0.33	1.33	4.33	0.33
5	0.33	0.67	3.67	0.33
6	0.00	2.33	6.00	0.33
7	2.00	7.67	30.00	0.67
Total	100.00	100.00	100.00	100.00

Table 5: Consistency of Updating Rules

Notes. The table reports how often participants use specific updating rules. Since participants complete seven updating tasks, they can apply each rule between 0 and 7 times (Column "Nr. Consistent Observations"). We allow for a distance of at most 2 p.p. between the prediction and the reported guess. The columns display the distribution of frequencies for Bayesian updating, one-stage updating, model selection, and reporting the prior. For example, the column "Bayesian" shows the share of participants who report guesses that correspond to Bayesian updating in 0, 1, 2, 3, 4, 5, 6, and 7 tasks. Note that in some cases, the predictions for these updating rules overlap: Bayesian and one-stage overlap in 2.33% of all observations, prior and one-stage in 1.86% of all observations, and prior and model selection in 3.05% of all observations. In such cases, the observation counts for both rules. No participant can have a majority of consistent guesses for multiple updating rules.

60% of participants when requiring a majority of guesses to follow one updating rule. Hence, we conclude that many participants consistently use a single updating rule.

In the appendix, we report a series of robustness tests from which we draw similar conclusions. Appendix Table A6 requires an exact match between the rule and the guess (instead of allowing for a distance of at most 2 percentage points) and Table A7 uses alternative approaches to classify participants based on average or median distances between their guesses and the predictions of the different updating rules. Moreover, we do not find that the updating rules participants use systematically change with experience.[27]

The observed pattern in Table 5 aligns with findings from the previous sections. While some participants put positive weight on both models by consistently following Bayesian and one-stage updating across tasks, the most common updating rule that is consistently applied corresponds to model selection, with 36% of participants using this rule in at least 6 out of 7 updating tasks. We then investigate the criteria these participants use for model selection. Figure 5 illustrates how frequently participants who select a model for all 7 tasks choose the best-fitting model. The figure reveals that a majority of these participants (51.11%) select the best-fitting model in at least 6 out of 7 tasks, while only 17.78% select the worst-fitting model the majority of the time. No participant selects the worst-fitting model in all the tasks. Appendix Figure A4 supports this finding by conducting

[27]Appendix Table A8 reports the joint distribution of used updating rules when comparing guessing the first and the last (before the repeated model pairs) tasks. This table reveals no systematic patterns, indicating no systematic change over time in the updating rules participants use.

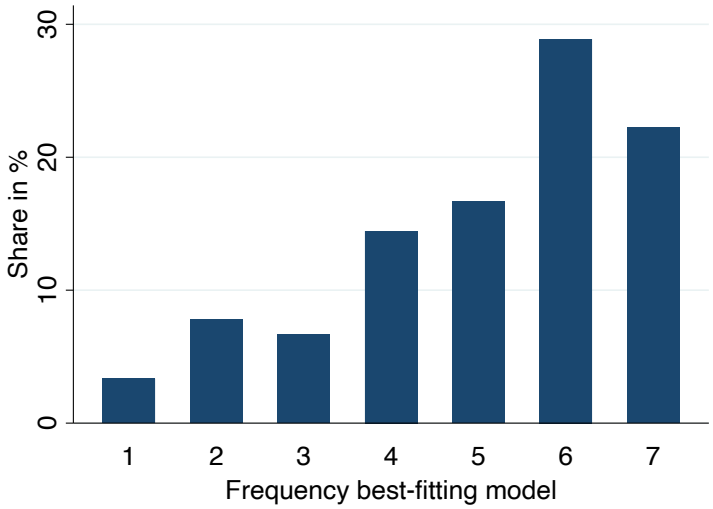


Figure 5: Errors in Model Selection

Notes. The figure shows, among the 90 participants who selected a model in all 7 tasks (see Table 5), how frequently they selected the best-fitting model rather than the worst-fitting one.

the same analysis for participants who selected a model in 6 out of 7 updating tasks. This systematic tendency to select the best-fitting model offers further support for the maximum likelihood criterion over the dogmatic and informativeness criteria. Appendix Figure A5 reports how frequently participants selected the most informative model and reveals that only a few participants consistently follow this criterion. Taken together, these findings reinforce the interpretation, discussed in the previous section, that model selection via maximum likelihood generally describes well which model participants pick, though errors in evaluating model fit can sometimes lead to the selection of the worst-fitting model.

4.5 Other Updating Rules

So far, we have focused on three main updating rules: Bayesian updating, one-stage updating, and model selection. These rules collectively explain most of our data: 72% of reported guesses align with these rules, allowing for a margin of error of at most 2 percentage points (Table 2). Moreover, participants display consistency in their use of these rules, with 53% applying the same rule for at least 5 out of 7 guesses (Table 5). Given that these rules account for a large share of the reported guesses, there is little room to explore the role of additional rules. Nonetheless, a small group of participants may consistently use alternative updating rules. In this section, we summarize our findings on these alternative updating rules and provide further details in Appendix D.

Less Extreme Forms of Over- and Underinference Model selection via maximum likelihood and one-stage updating reflect extreme over- and underinference about the models, respectively. Some participants might instead engage in less extreme versions of such biased inference about the models. In Appendix D.1, we present a theoretical

framework capturing such rules and apply it to our data. This analysis reveals that some participants consistently use less extreme forms of over- or underinference but occasionally make errors in identifying the best-fitting model. Incorporating these updating rules enables us to classify an additional 16% of participants when requiring consistency across at least 5 out of 7 guesses, increasing the total share of classified participants to 70%.

Signal-independent Model Weights One-stage updating and dogmatic model selection assign model weights that do not depend on the signal. In Appendix D.2, we study the broader class of updating rules sharing this property. However, we find no evidence that any other signal-independent updating rules are consistently applied in our data.

Aggregating Competing Models into a Single Model Instead of aggregating the predictions of the different models, individuals could first combine the competing models into a single compound model and then use this model to update their beliefs based on the observed signal. For a Bayesian, these two approaches are equivalent if the compound model is derived using the prior over models. However, biases in updating with such a compound model result in deviations from the Bayesian benchmark that differ fundamentally from those arising from the biases in model weighting discussed in Section 2. Most importantly, updating based on a compound model predicts posteriors to be Bayes-consistent, even if the updating process or the compounding procedure is biased. However, we find that half of the reported vectors of posteriors are Bayes-inconsistent. Appendix D.3 provides further evidence against the consistent application of such rules that are based on a compound model.

Use of Multiple Rules Participants might alternate between different updating rules, such as using one-stage updating in some instances and Bayesian updating in others. However, as shown in Appendix D.4, participants rarely use multiple updating rules, and no systematic patterns emerge in their combinations.

Hence, in addition to one-stage updating, Bayesian updating, and model selection, we find that some participants employ less extreme forms of over- or underinference about the models. However, our analysis finds no evidence that any other rules are applied frequently or consistently in our data.

4.6 Determinants of Model Selection versus Model Weighting

Having established that many participants consistently use specific updating rules, we now investigate whether different groups of participants employ different rules. Our main focus is on understanding the characteristics of individuals who engage in model selection rather than consistently assigning positive weight to both models, as these individuals

Dependent Variable	Model Selection								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.180***								0.147**
	(0.069)								(0.069)
Age		-0.005							-0.005*
		(0.003)							(0.003)
Education			-0.063						-0.009
			(0.047)						(0.042)
Very Liberal				-0.050					-0.067
				(0.077)					(0.071)
Very Conservative					0.205**				0.091
					(0.083)				(0.088)
Cognitive Reflection Test						-0.132***			-0.109***
						(0.025)			(0.026)
REI Rationality Scale							-0.161***		-0.107*
							(0.056)		(0.056)
REI Experientiality Scale								0.054	0.002
								(0.045)	(0.041)
Observations	162	162	162	162	162	162	162	162	162
R^2	0.042	0.015	0.012	0.003	0.016	0.114	0.041	0.011	0.185

Table 6: Explaining Model Selection

Notes. The table shows coefficient estimates from linear regressions of an indicator of model selection on different explanatory variables. The sample consists of the 162 participants who consistently used one updating rule (Bayesian, one-stage, model selection, prior) in at least 5 out of 7 tasks, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). "Female," "Very Liberal," and "Very Conservative" are indicator variables, "Age" is in years (19 and 70), "Education" is from 1 = "High school" to 4 = "PhD or higher," "Cognitive Reflection Test" are scores from 0 to 3, and Rational Experiental Inventory (REI) rationality and experientiality scales are from 1 to 5. Heteroskedasticity-robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

are particularly vulnerable to persuasion using models.[28]

Table 6 presents the results of regressing a dummy variable for model selection on various participant characteristics. Our analysis focuses on the 162 participants who consistently use one of the updating rules in at least 5 out of 7 tasks. Appendix Tables A9, A10, A11, and A12 demonstrate that these results are robust when focusing on participants who used one rule in 4, 6, or 7 tasks, or when considering the entire sample and counting the number of times a participant selected a model.

We begin our analysis by considering the main demographic characteristics of our participants. Specifications (1) to (3) provide coefficient estimates for gender, age, and

[28] Appendix Figure A6 presents the characteristics of participants who apply different updating rules, also distinguishing between the two main rules that put positive weight on both models—Bayesian updating and one-stage updating—and model selection.

education. We find that women are relatively more likely to select models rather than weight them. Next, we examine self-reported political attitudes. Model selection has been theoretically linked to holding particularly extreme views. Hence, since the relationship between model selection and political attitudes could be nonlinear—where individuals with more extreme positions are more likely to select models—we include dummies for extreme political positions, "Very Liberal" and "Very Conservative". Specifications (5) and (6) confirm that being very conservative is associated with more frequent model selection, while no statistically significant association is found for strongly liberal views.

Biases in belief updating have also been linked to cognitive abilities (Oechssler et al., 2009; Hoppe and Kusterer, 2011; Augenblick et al., 2025b; Enke and Graeber, 2023; Oprea, 2025).[29] We collect different psychological assessments of cognitive ability and thinking styles at the end of our study. First, we consider the role of the Cognitive Reflective Test (CRT), which measures the tendency to override incorrect intuitive responses (Frederick, 2005). Specification (6) shows that participants with high CRT scores are less likely to select models. This can be explained by noting that individuals with higher CRT scores tend to engage in more deliberative and analytical processing of information, making them more likely to weight competing models rather than selecting a single one. Second, we consider the Rational-Experiential Inventory (REI), which assesses participants' thinking styles (Keaton, 2017). This scale measures the extent to which participants engage in two modes of thinking: intuitive thinking (Experientiality Scale) and logical thinking (Rationality Scale). Specifications (7) and (8) indicate that participants with higher scores on the REI Rationality Scale are less likely to select models, while there is no significant effect for the Experientiality Scale. This further suggests that individuals who favor a methodical and rational approach to processing information are more inclined to weight multiple models rather than selecting a single one.

Finally, in Specification (9), we assess all these factors in a single regression. We find that CRT scores and REI remain significant predictors when all factors are included. However, the association with being very conservative is no longer statistically significant when controlling for all these measures. Hence, differences in updating based on political position seem to be shaped by underlying differences in thinking style.

These findings suggest that cognitive thinking styles, as measured by the CRT and REI, might be particularly relevant in determining whether a participant chooses to weight multiple models or select a single one, and, hence, how vulnerable to persuasion using models this individual can be.

[29] As discussed in Appendix E, the updating rules involve a trade-off between precision and response time. Participants who are classified as consistently engaging in model selection or one-stage updating report beliefs that deviate from the Bayesian benchmark, but also exhibit substantially shorter response times. Specifically, the median response time for Bayesian updating is more than three times higher than that for model selection. Hence, model selection may reduce cognitive burden, making it particularly prevalent among individuals with more limited cognitive capacity.

4.7 Under- and Overinference about the State

So far, our analysis has focused on biases in model weighting. In contrast, the existing literature on belief updating typically examines over- and underinference regarding the state of the world—and not about models—because it assumes a single data-generating process. In this section, we connect these two perspectives and discuss how biases in model weighting relate to biases in beliefs about the state of the world.

In our data, 3.05% of observations correspond to Bayesian updating, 45.62% reflect overinference about the state, 17.86% reflect underinference about the state, and 33.48% involve updating in the opposite direction of what Bayesian updating about the state would predict.

Drawing a general conclusion about overall over- or underinference about the state in our data is challenging because a substantial fraction of the observations involve updating in the wrong direction. If these observations are excluded, as sometimes done in the literature in settings with a single model, the data would suggest overinference about the state. In contrast, including them leads to underinference about the state. Indeed, using the standard approach in the literature based on the model introduced by Grether (1980), we estimate $\alpha_s = 0.686$ (N= 1,992, t = 16.48, 95% CI = [0.604, 0.768]), which would be interpreted as substantial underinference.[30] This finding is consistent with Liang (2025).[31]

Appendix Table A13 illustrates that the relationship between the degree of inference about models and inference about the state is not straightforward.[32] First, overinference about models often leads to updates that would be classified as underinference about the state, or, most frequently in our data, updating in the wrong direction. Indeed, for conflicting models, model selection via maximum likelihood—the extreme case of overinference about models—leads to overinference about the state for one signal realization, but leads for the other realization to either updating in the wrong direction or, when one model is uninformative, underinference about the state of the world. Second, underinference about models frequently results in beliefs that would be classified as overinference about the state. As a result, only 32.15% of observations show a match between the

[30]Specifically, we regress the logit of posterior beliefs on the logit of the Bayesian posterior, without a constant. The estimated coefficient, α_s, captures the degree of overinference ($\alpha_s > 1$) or underinference ($\alpha_s < 1$). Observations with extreme beliefs equal to 0 or 100 (108 out of 2,100) are excluded because a logit cannot be computed; results are robust to retaining these observations and coding them as 99 and 1 instead. Note that this approach is designed to study belief updating when individuals entertain a single model and does not fit our data well (see Section 4.5 and Appendix D.3.1 for details). We report these results solely for comparison with the existing literature.

[31]Liang (2025) documents underreaction to uncertain information, which in our setting corresponds to beliefs about the state being less responsive to signals when individuals encounter multiple models rather than a single one. Our data are consistent with this finding: using observations from Part 1 where participants face a single model, we estimate $\alpha_s = 0.816$ (N = 577, t = 19.69, 95% CI = [0.735, 0.898]), indicating a significantly lower degree of underinference (t = 2.19, p = 0.030).

[32]Comparing Appendix Figures D1 and D6 highlights how the same vectors of posteriors given the same model pair can imply very different classifications of bias depending on whether one focuses on model weighting or on inference about the state.

classification of biases for models and for states (e.g., overinference about models leads to overinference about the state).

5 Replication and Generalizability

So far, we have focused on updating environments characterized by (i) an objective prior distribution over models, (ii) readily available model predictions, (iii) the presence of two models, and (iv) conflicting models—that is, models that make opposing predictions about the state of the world.

We chose this environment as a starting point because it captures important features of many real-world settings and allows us to reliably measure biases in model weighting within a clean and simple experimental design. The results presented so far indicate that many individuals use model selection to update their beliefs. In this section, we present results from two additional studies that independently vary these four design features to address additional questions relevant to these alternative contexts and to assess the generalizability of our findings.

To ensure the comparability of our results, each study incorporates a *Baseline* condition that replicates the initial design; this serves as a benchmark against which we assess behavior in new treatments that modify key aspects of the environment. While the additional studies closely follow the design described in Section 3, they use a subset of the original guessing tasks, specifically Model Pair 2 and Model Pair 3, as described in Table 1 and Figure 4. Participants complete both updating tasks, and in the third study, they also repeat one at random, allowing us to recover vectors of posterior beliefs. Importantly, the data from the *Baseline* conditions replicate the findings from our main study, confirming the robustness of our results.[33]

The second study was conducted on Prolific in February 2025 with 592 participants across three treatments; the third was conducted on Prolific in December 2025 with 586 participants across three treatments. Appendix F provides further details on the experimental designs and additional results.[34]

5.1 No Priors over Models

We first examine an environment in which participants are not provided with an exogenous prior distribution over models. This setting aligns with the literature on ambiguous beliefs, which does not specify model priors and also considers model selection via maximum likelihood. In the *No-Prior* condition, we study whether model selection remains a

[33]See Tables 7 and 9. As a benchmark, Appendix Table F2 reports the shares of updating rules for Model Pair 2 and Model Pair 3 in the initial study.

[34]The pre-registration plans are available at `https://aspredicted.org/wqtd-68zm.pdf` and `https://aspredicted.org/78jp45.pdf`.

prevalent updating rule in contexts where participants are not provided with an exogenous prior over the models.

Table 7 reports the classification of guesses consistent with Bayesian updating, one-stage updating, and selecting either the best- or worst-fitting model.[35] We find that model selection is as prevalent in the *No-Prior* condition as in the *Baseline* condition; allowing for small deviations of up to 2 percentage points, 53.7% versus 55.5% of reported guesses correspond to selecting a single model. As in the *Baseline* condition, selecting the best-fitting model is the most common updating rule (44%), followed by one-stage updating (13%). None of the shares in the *No-Prior* condition differ statistically or meaningfully from those in the *Baseline* condition (Appendix Tables F5 and F6).

Type of Guess	Baseline		No-Prior		Click	
	%	95%-CI	%	95%-CI	%	95%-CI
			Exact			
Bayesian	0.75	[0.00, 1.59]	0.91	[0.00,2.01]	2.03	[0.54, 3.53]
One-stage	8.71	[5.13, 12.28]	8.90	[5.41, 12.40]	11.92	[7.57, 16.27]
Best-fitting	36.07	[30.28, 41.85]	36.99	[31.61, 42.36]	38.95	[32.40, 45.51]
Worst-fitting	10.70	[7.37, 14.02]	7.76	[5.10, 10.42]	9.88	[6.19, 13.58]
Within	31.09	[25.55, 36.64]	29.91	[24.75, 35.06]	25.29	[19.68, 30.90]
Outside	12.69	[8.81, 16.56]	15.53	[11.89, 19.16]	11.92	[8.22, 15.61]
Total	100.00		100.00		100.00	
			Within 2 p.p.			
Bayesian	3.73	[1.90, 5.56]	4.57	[2.44, 6.69]	4.94	[2.55, 7.34]
One-stage	12.19	[8.10, 16.28]	13.47	[9.46, 17.48]	15.99	[11.19, 20.78]
Best-fitting	44.03	[38.02, 50.04]	43.61	[38.04, 49.18]	44.48	[37.79, 51.16]
Worst-fitting	11.44	[7.98, 14.90]	10.05	[7.01, 13.08]	11.34	[7.51, 15.16]
Total	71.39		71.70		76.75	

Table 7: Classification of Guesses by Treatment (Second Study)

Notes. Section "Exact" reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, and worst-fitting) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise). Section "Within 2 p.p." reports the shares of guesses that fall within 2 p.p. around each point prediction. The table reports results for each treatment (*Baseline*, *No-Prior*, *Click*). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We pool the data from both updating tasks.

[35]Note that while model selection remains well-defined in the *No-Prior* condition, identifying Bayesian and one-stage updating requires assumptions about subjective model priors. As pre-registered, Table 7 focuses on Bayesian and one-stage updating with uniform prior over models. However, participants could use Bayesian or one-stage updating based on subjective model priors other than the uniform prior. However, two findings suggest little room for such updating. First, we find no evidence of participants frequently assigning weights to models beyond those discussed in Table 7. Moreover, when we consider Bayesian and one-stage updating assuming equal prior over models, we find that the prevalence of these updating rules in the *No-Prior* condition matches their prevalence in the *Baseline* condition.

5.2 Demand for Model Predictions

Second, we examine a setting where model predictions are not immediately available, requiring participants to exert effort to access them. While providing model predictions reflects many real-world settings where individuals encounter a model alongside its prediction, in other cases, such predictions may not always be readily available and involve cognitive, attention, or search costs to access them. In the *Click* condition, participants can obtain each prediction by clicking a corresponding button five times in a row ("R" for the first model and "B" for the second model). This approach introduces real effort costs in a highly controlled manner (e.g., Ariely et al., 2009; DellaVigna and Pope, 2018), capturing any potential costs associated with accessing model predictions. This condition allows us to examine whether participants actively seek out model predictions at a small cost and whether model selection remains prevalent in settings where model predictions are less salient and require deliberate effort to access them.

We start by examining the demand for model predictions. Table 8 presents the frequencies with which participants revealed zero, one, or both model predictions. Participants are generally willing to take a deliberate, costly action to access model predictions. In only 6% of all updating tasks, participants proceed without acquiring any predictions, while in 10% of tasks participants reveal the prediction of one model, and in 84% of tasks they reveal predictions for both models. Interestingly, when participants reveal only one model prediction, it is more often the prediction of the best-fitting model (6.40% vs. 3.20%; $t = 2.12, p = 0.046$).

Nr. Predictions Revealed	%	95%-CI
Zero	6.10	[2.73, 9.48]
One, Total	9.59	[5.58, 13.61]
One, Best-fitting	6.40	[3.39, 9.40]
One, Worst-fitting	3.20	[1.18, 5.22]
Both	84.01	[78.95, 89.08]

Table 8: Demand for Model Predictions

Notes This table reports the frequencies with which participants revealed zero, one, or both model predictions in the *Click* condition (N=344). For revealing one prediction, the table gives the overall frequency (row "One, Total"), as well as the frequency in which only the predictions of the best-fitting and only the worst-fitting models are revealed (rows "One, Best-fitting" and "One, Worst-fitting"). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We pool the data from both updating tasks.

Turning to updating patterns, we find that the *Click* condition closely mirrors the *Baseline*.[36] Table 7 confirms that there are no substantial differences in updating rules employed; in particular, selecting the best-fitting model remains very prevalent. None of the

[36]We cannot determine whether participants who do not reveal any model predictions are following one of the discussed updating rules or suffer from biases in deriving model predictions instead. However, the high rate at which participants revealed the predictions ensures comparability between the *Click* and *Baseline* conditions.

shares of guesses from the *Click* condition are statistically significantly or meaningfully different from those of *Baseline* (Appendix Tables F5 and F6).

5.3 Non-Conflicting Models

Next, we examine belief updating in the presence of non-conflicting models, that is, models that agree on the direction of the update but differ in the magnitude of their predictions. While our main focus on conflicting models was motivated by their relevance for polarization, this new environment serves a useful comparative purpose. Existing theories do not predict that updating should depend on whether models are conflicting; however, individuals could be less willing to entertain two fundamentally opposing models than two non-conflicting ones. This condition allows us to test this possibility.

To ensure comparability with the *Baseline* condition, we construct the *Non-Conflicting* condition by replacing Model Pair 2 and Model Pair 3 with non-conflicting counterparts. Specifically, we select these new model pairs, presented in Appendix Table F1, to induce the same Bayesian posteriors as their conflicting counterparts. This design keeps the informational content constant across conditions.

Table 9 shows no substantial differences in updating rules relative to the *Baseline* condition.[37] Model selection—and especially selection of the best-fitting model—remains prevalent, and none of the shares differ statistically or meaningfully from the *Baseline* condition (Appendix Table F10).

From a theoretical perspective, the distinction between conflicting and non-conflicting models is decisive for the emergence of Bayes-inconsistencies. With non-conflicting models, even extreme biases in model weighting—such as assigning full weight to the best-fitting model—translate to beliefs that are Bayes-consistent; Bayes-inconsistencies are predicted to arise only when models are conflicting. Hence, the *Non-Conflicting* condition provides a key benchmark: if the Bayes-inconsistencies observed in the main study with conflicting models were driven by noise, they should still be common in a setting with non-conflicting models. If, instead, these are largely absent, it would indicate that the inconsistencies with conflicting models are primarily driven by biases in model weighting.

Table 10 shows that Bayes-inconsistencies are rare in the *Non-Conflicting* condition, accounting for fewer than 10% of observations. In contrast, in the *Baseline* condition, we replicate the finding from the initial study: half of the observations are Bayes-inconsistent, and most are in the direction of the best-fitting model. We estimate that Bayes-inconsistencies are 45.6 percentage points less frequent in the *Non-Conflicting* condition (OLS with heteroskedasticity-robust standard errors: $t = -7.83$, $p < 0.001$).

[37] In our third study, our classification focuses on the exact match between the guess and the predictions of the different rules. Allowing for deviations of up to 2 percentage points results in nearly 20% of observations in both the *Non-Conflicting* and *Three-Models* conditions being consistent with multiple rules, making the interpretation of the results challenging. Appendix Table F7 provides these results.

Type of Guess	Baseline		Non-Conflicting		Three-Models	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	2.45	[0.57, 4.33]	2.13	[0.67, 3.58]	7.04	[4.73, 9.36]
One-stage	9.31	[5.88, 12.75]	13.83	[9.57, 18.09]	15.29	[10.89, 19.70]
Best-fitting	42.48	[36.79, 48.18]	38.48	[32.62, 44.33]	27.66	[22.52, 32.81]
Worst-fitting	6.54	[4.03, 9.04]	7.80	[5.28, 10.32]	3.95	[1.91, 6.00]
Middle Model					10.31	[7.21, 13.41]
Within	28.92	[23.92, 33.92]	25.89	[21.07, 30.70]	31.62	[26.34, 36.89]
Outside	10.29	[7.41, 13.17]	11.88	[8.18, 15.58]	10.14	[7.23, 13.04]
Total	100.00		100.00		100.00	

Table 9: Classification of Guesses by Treatment, Exact Match (Third Study)

Notes. The table reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, worst-fitting, and middle model) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise) for each treatment (*Baseline*, *Non-Conflicting*, *Three-Models*). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We pool the data from all updating tasks. Note that, in *Three-Models*, 6.01% of reported guesses are classified both as Bayesian and middle model. These duplicate observations are not included in the total.

Consistency	Baseline		Non-Conflicting		Three-Models	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayes-consistent	45.00	[35.08, 54.92]	90.63	[84.69, 96.56]	53.76	[43.44, 64.09]
Bayes-inconsistent (best-f.)	49.00	[39.03, 58.97]	5.21	[0.68, 9.73]	38.71	[28.62, 48.80]
Bayes-inconsistent (worst-f.)	6.00	[1.26, 10.74]	4.17	[0.10, 8.24]	7.53	[2.06, 12.99]
Total	100.00		100.00		100.00	

Table 10: Bayes-inconsistencies (Third Study)

Notes. This table describes Bayes-consistent and Bayes-inconsistent vectors of beliefs in the *Baseline* (N=100), *Non-Conflicting* (N=96) and *Three-Model*s (N=93) conditions. Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using heteroskedasticity-robust standard errors. For the *Baseline* and *Three-Models* conditions, the row "Bayes-inconsistent (best-f.)" reports the share of Bayes-inconsistent observations that are in the direction of the best-fitting model, relative to all observations; such inconsistencies are predicted by biases in model weighting that systematically place disproportionate weight on the best-fitting model. The row "Bayes-inconsistent (worst-f.)" reports the share of Bayes-inconsistent observations that are in the direction of the worst-fitting model, relative to all observations. For non-conflicting models, neither consistently selecting the best-fitting nor the worst-fitting model generates Bayes-inconsistencies. We construct these categories to mirror the directions of the best- and worst-fitting models in the corresponding conflicting model pairs.

5.4 Three Models

Finally, we study updating in environments with more than two models to assess whether model selection remains prevalent. In the *Three-Models* condition, we add a third model to the original model pairs in the *Baseline* condition (see Appendix Table F1). For better comparability, this additional model (the "middle model") is constructed to have a fit level between the ones of the other models, so that the ranking of the best- and worst-fitting models remains unchanged.

A key feature of our design is that it makes the correct Bayesian prediction as salient as the model predictions in one of the updating tasks. For Model Pair 2, the middle model is the compound of the two original models. Given the uniform prior over the models, the prediction of this middle model corresponds to the Bayesian posterior for this task.[38] This allows us to make the Bayesian prediction salient and to study the persuasiveness of a model that provides the Bayesian prediction. Importantly, the fit of such a compound model that provides the Bayesian prediction is a convex combination of the fits of the other models, meaning it is never the best-fitting one. As a result, biases in model weighting that place disproportionate weight on the best-fitting model predict that such a model will not be persuasive.

As shown in Table 9, model selection continues to play a central role in the *Three-Models* condition (49.0% in *Baseline* vs. 41.9% in *Three-models*; $p = 0.11$, Appendix Table F10). Most importantly, selecting the best-fitting model remains the most frequently selected updating rule.[39]

Although the middle model is selected in approximately 10% of observations, the best-fitting model is chosen nearly three times as often. Appendix Table F8 shows that this pattern persists even in the tasks where selecting the middle model coincides with Bayesian updating (and thus, where some of these choices likely reflect Bayesian updating rather than selecting the middle model). These results show that providing the Bayesian prediction is not sufficient to make a model persuasive. Moreover, they demonstrate that model selection via maximum likelihood remains the modal updating rule even when the Bayesian prediction is as salient as that of the best-fitting model.

Appendix Table F10 also highlights a few differences compared to an environment with two models. First, conditional on model selection, the best-fitting model is selected slightly less frequently in the *Three-Models* condition ($p < 0.01$), while one-stage updating becomes more common ($p < 0.04$). Second, Bayesian updating also appears more frequent

[38] For Model Pair 3, the compound model cannot be represented with urns of six balls. Instead, we use a middle model that approximates the corresponding compound model, yielding a prediction strictly between the two other model predictions. This provides a benchmark for how often such an intermediate model is chosen when it does not coincide with the Bayesian prediction.

[39] In line with the prevalence of model selection via maximum likelihood, Table 10 shows that Bayes-inconsistencies remain frequent in the *Three-Models* condition; we cannot reject that these inconsistencies occur less often than in the *Baseline* condition (OLS with heteroskedasticity-robust standard errors: $t = -1.22$, $p = 0.225$).

for three models; however, this is driven by the overlap between the Bayesian prediction and that of the middle model. Indeed, when restricting the analysis to the tasks where no such overlap exists, we observe no differences in the share of observations that correspond to Bayesian updating between the two conditions (Appendix Table F12).

Taken together, these results demonstrate that model selection via maximum likelihood continues to play a central role in environments with three models, although we also find some general differences in updating patterns.[40]

6 Discussion and Conclusion

This paper studies how individuals update their beliefs when confronted with competing data-generating processes, or models, that could explain the data and provide conflicting predictions. We design and implement an experimental study to identify the weights individuals assign to different models when updating their beliefs in such situations.

We offer three key insights. First, we provide strong evidence supporting the most common assumption in the literature on model persuasion: model selection via maximum likelihood. This approach assumes that individuals place full weight on a single model, specifically, the one that best fits the data. Our findings show that model selection via maximum likelihood is the most frequently applied updating rule across different updating environments, including those with and without a specified model prior, when people encounter conflicting or non-conflicting models, and when the correct Bayesian prediction is as salient as the prediction of the best-fitting model. Moreover, participants apply this rule consistently across updating tasks. These results not only validate the assumption of theoretical papers about model persuasion but also hold important implications for policy-oriented research. In particular, understanding how people aggregate the predictions from competing models can shed light on the cognitive processes that drive polarization. While assigning positive weights to multiple models might promote more nuanced perspectives, the tendency to select only one model could contribute to the rise of extreme beliefs and polarized views (Izzo et al., 2023; Aina, 2025; Schwartzstein and Sunderam, 2024).

Second, while model selection via maximum likelihood describes well the belief updating of many participants, our study uncovers additional belief-updating patterns that inform future research, especially in developing more descriptively accurate models to assess policy impact and welfare considerations. On one hand, we observe that individuals often make errors when applying model selection criteria, at times selecting the worst-fitting

[40]When faced with three models, participants might simplify the problem by first eliminating one model from consideration and then applying either Bayesian or one-stage updating on the remaining pair. We find that such behavior is uncommon, accounting for 5.67% of observations in the *Three-Models* condition (split evenly between Bayesian and one-stage updating). Among the observations consistent with these rules, the model excluded from the subset is typically the middle (45%) or worst-fitting (36%) model, rather than the best-fitting model (18%). This suggests that even when participants reduce the model space, their exclusion criteria are guided by model fit.

model instead of the best-fitting one. These mistakes differ from the usual approach to modeling errors, which involves frequent but minor deviations from the predictions of the underlying updating rule. In contrast, errors in model selection are of large magnitude. Considering the examples discussed in the introduction, such errors could lead to fundamentally different conclusions about future asset returns, the legitimate winner of an election, or vaccine safety. On the other hand, we find that some participants consistently apply updating rules other than selecting the best-fitting model. In particular, one-stage updating emerges as the second most frequently applied rule in our study, and we believe it deserves attention for future research. From the perspective of over- and underinference about the models, one-stage updating stands as the opposite extreme of model selection via maximum likelihood: While model selection via maximum likelihood represents an extreme form of overinference about the models, one-stage updating represents an extreme form of underinference. A direct implication of errors in applying model selection criteria and one-stage updating is that the distribution of beliefs about the state of the world is multimodal, with peaks at the predictions of the best-fitting model, the worst-fitting model, and the one-stage updating. Consequently, average beliefs fail to adequately represent the population's belief distributions, as in Bordalo et al. (2026).

Third, our data shows the systematic emergence of Bayes-inconsistencies: when confronted with conflicting models, individuals become more confident about a certain state of the world regardless of the observed signal. Bayes-inconsistencies are a central prediction of certain biases in model weighting, particularly those that lead to strong overinference about the models. This finding demonstrates the potential impact of persuasion using models. Bayes-inconsistencies contrast with the constraint the persuader faces in classic models of persuasion, such as Bayesian persuasion. It shows how it is possible to persuade individuals to adopt Bayes-inconsistent beliefs by presenting them with conflicting models, as proposed by Aina (2025). For instance, a politician might convince voters that she won an election regardless of the reported outcome by introducing conflicting narratives about the election system. Voters could selectively adopt these narratives depending on the result, concluding that the system is fair if she wins but rigged if she loses. Analogously, a financial advisor could convince investors to always invest in certain assets by providing conflicting ways of interpreting past financial data.

The observation of frequent Bayes-inconsistencies also highlights that biases in model weighting can have fundamentally different implications than those associated with entertaining a single model. There are no theoretical reasons to expect this type of inconsistent beliefs when individuals update their beliefs based on a single model, even when exhibiting over- or underreaction to the new information. These observations are also important for researchers studying belief updating, as it is often challenging to determine whether individuals consider multiple models and, if so, which models they consider. We provide two valuable insights on this perspective. First, studying biases in belief updating in a single-model framework, when individuals may actually entertain multiple models, can lead to inaccurate conclusions. For example, frequent Bayes-inconsistencies would

be attributed to belief updates moving in the wrong direction and, consequently, to low data quality. Therefore, it is important to carefully consider the possibility that subjects entertain multiple models when analyzing belief data. Second, even when models are not directly observable, one can assess whether subjects entertain multiple conflicting models by testing for frequent occurrences of Bayes-inconsistencies. Such inconsistencies serve as a clear indicator that participants are considering multiple models but exhibit biases in how they weight them.

Several compelling questions remain to be addressed. For instance, we deliberately focus on a setting where individuals have no personal stake in the outcomes. Introducing stakes could lead to additional motives; for example, individuals might be inclined to select models that offer the most optimistic predictions regarding their potential rewards. Moreover, we limit our analysis to a context in which individuals update their beliefs in response to only one signal. This approach leaves the dynamic aspects of this process unexplored, raising questions about how model selection might evolve in response to multiple signals over time: would individuals re-evaluate models based on cumulative signals, assign more importance to the more recent signal, or adhere to the first model selected? This paper establishes a foundation to understand belief updating in these settings, paving the way for further research about decision-making when more than one model informs our interpretation of what we observe.

References

Aina, Chiara (2025) "Tailored Stories."

Aina, Chiara, Andrea Amelio, and Katharina Brütt (2025) "Contingent Belief Updating."

Ambuehl, Sandro and Heidi C Thysen (2024) "Choosing Between Causal Interpretations: An Experimental Study."

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart (2023) "Narratives about the Macroeconomy."

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2022) "Subjective models of the macroeconomy: Evidence from experts and representative samples," *Review of Economic Studies*, 89 (6), 2958–2991.

Ariely, Dan, Anat Bracha, and Stephan Meier (2009) "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially," *American Economic Review*, 99 (1), 544–555.

Augenblick, Ned, Matthew Backus, Andrew T Little, and Don A Moore (2025a) "Assumptions, Disagreement, and Overprecision: Theory and Evidence."

Augenblick, Ned, Eben Lazarus, and Michael Thaler (2025b) "Overinference from Weak Signals and Underinference from Strong Signals," *Quarterly Journal of Economics*, 140 (1), 335–401.

Ba, Cuimin (2024) "Robust Misspecified Models and Paradigm Shifts."

Ba, Cuimin, J Aislinn Bohren, and Alex Imas (2025) "Over-and Underreaction to Information."

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998) "A model of investor sentiment," *Journal of Financial Economics*, 49 (3), 307–343.

Barron, Kai and Tilman Fries (2024) "Narrative Persuasion: A Brief Introduction."

——— (2025) "Narrative Persuasion."

Bauch, Gerrit and Manuel Foerster (2024) "Strategic communication of narratives."

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

Bohren, J Aislinn and Daniel N Hauser (2024) "Behavioral Foundations of Model Misspecification."

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer (2026) "How people use statistics," *Review of Economic Studies*, 93 (1), 250–285.

Bordalo, Pedro, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer (2025) "A cognitive theory of reasoning and choice."

Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott (2023) "Opinions as facts," *Review of Economic Studies*, 90 (4), 1832–1864.

Camerer, Colin (1995) "Individual decision making," *The Handbook of Experimental Economics*, 1, 587–704.

Çelebi, Can, Christine Exley, Sören Harrs, Hannu Kivimaki, Marta Serra-Garcia, and Jeffrey Yusof (2025) "Mission Possible: The Collection of High-Quality Data."

Chater, Nick and Mike Oaksford (2008) *The probabilistic mind: Prospects for Bayesian cognitive science*: Oxford University Press, USA.

Cokely, Edward T, Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-Retamero (2012) "Measuring risk literacy: The Berlin numeracy test," *Judgment and Decision making*, 7 (1), 25–47.

Danz, David, Lise Vesterlund, and Alistair J Wilson (2022) "Belief elicitation and behavioral incentive compatibility," *American Economic Review*, 112 (9), 2851–83.

DellaVigna, Stefano and Devin Pope (2018) "What motivates effort? Evidence and expert forecasts," *Review of Economic Studies*, 85 (2), 1029–1069.

Dempster, Arthur P (1967) "Upper and lower probability inferences based on a sample from a finite univariate population," *Biometrika*, 54 (3-4), 515–528.

Douven, Igor and Jonah N Schupbach (2015a) "Probabilistic alternatives to Bayesianism: the case of explanationism," *Frontiers in Psychology*, 6, 459.

——— (2015b) "The role of explanatory considerations in updating," *Cognition*, 142, 299–311.

Edwards, Ward (1968) "Conservatism in human information processing," *Formal representation of human judgment*.

Eliaz, Kfir, Simone Galperti, and Ran Spiegler (2025) "False narratives and political mobilization," *Journal of the European Economic Association*, 23 (3), 983–1027.

Eliaz, Kfir and Ran Spiegler (2020) "A model of competing narratives," *American Economic Review*, 110 (12), 3786–3816.

——— (2024) "News Media as Suppliers of Narratives (and Information)."

Eliaz, Kfir, Ran Spiegler, and Yair Weiss (2021) "Cheating with models," *American Economic Review: Insights*, 3 (4), 417–434.

Enke, Benjamin and Thomas Graeber (2023) "Cognitive uncertainty," *Quarterly Journal of Economics*, 138 (4), 2021–2067.

Epstein, Larry G and Yoram Halevy (2024) "Hard-to-interpret signals," *Journal of the European Economic Association*, 22 (1), 393–427.

Fréchette, Guillaume, Sevgi Yuksel, and Emanuel Vespa (2024) "Extracting Models from Data Sets: An Experiment."

Frederick, Shane (2005) "Cognitive reflection and decision making," *Journal of Economic Perspectives*, 19 (4), 25–42.

Fryer, Roland G, Jr, Philipp Harms, and Matthew O Jackson (2019) "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization," *Journal of the European Economic Association*, 17 (5), 1470–1501.

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

Gilboa, Itzhak and David Schmeidler (1993) "Updating ambiguous beliefs," *Journal of Economic Theory*, 59 (1), 33–49.

Graeber, Thomas, Christopher Roth, and Constantin Schesch (2024a) "Explanations."

Graeber, Thomas, Christopher Roth, and Florian Zimmermann (2024b) "Stories, statistics, and memory," *Quarterly Journal of Economics*, qjae020.

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

Hoppe, Eva I and David J Kusterer (2011) "Behavioral biases and cognitive reflection," *Economics Letters*, 110 (2), 97–100.

Hossain, Tanjim and Ryo Okui (2013) "The binarized scoring rule," *Review of Economic Studies*, 80 (3), 984–1001.

Ichihashi, Shota and Delong Meng (2021) "The Design and Interpretation of Information."

Ispano, Alessandro (2025) "The perils of a coherent narrative," *Economic Theory*.

Izzo, Federica, Gregory J Martin, and Steven Callander (2023) "Ideological Competition," *American Journal of Political Science*, 67 (3), 687–700.

Jain, Atulya (2024) "Informing agents amidst biased narratives."

Kamenica, Emir and Matthew Gentzkow (2011) "Bayesian persuasion," *American Economic Review*, 101 (6), 2590–2615.

Keaton, Shaughan A (2017) "Rational-Experiential Inventory-40 (REI-40) (Pacini & Epstein, 1999)," *The sourcebook of listening research: Methodology and measures*, 530–536.

Kendall, Chad and Ryan Oprea (2024) "On the complexity of forming mental models," *Quantitative Economics*, 15 (1), 175–211.

Kendall, Chad W and Constantin Charles (2025) "Causal Narratives."

Levy, Gilat and Ronny Razin (2021) "A maximum likelihood approach to combining forecasts," *Theoretical Economics*, 16 (1), 49–71.

Liang, Yucheng (2025) "Learning from unknown information sources," *Management Science*, 71 (5), 3873–3890.

Lord, Charles G, Lee Ross, and Mark R Lepper (1979) "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.," *Journal of Personality and Social Psychology*, 37 (11), 2098.

Mullainathan, Sendhil (2002) "Thinking through categories."

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008) "Coarse thinking and persuasion," *Quarterly Journal of Economics*, 123 (2), 577–619.

Musolff, Robin and Florian Zimmermann (2025) "Model uncertainty."

Oechssler, Jörg, Andreas Roider, and Patrick W Schmitz (2009) "Cognitive abilities and behavioral biases," *Journal of Economic Behavior & Organization*, 72 (1), 147–152.

Oprea, Ryan (2025) "Complexity and Its Measurement."

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

Rabin, Matthew and Joel L Schrag (1999) "First impressions matter: A model of confirmatory bias," *Quarterly Journal of Economics*, 114 (1), 37–82.

Samuelson, Larry and Jakub Steiner (2024) "Constrained Data-Fitters."

Schwartzstein, Joshua and Adi Sunderam (2021) "Using models to persuade," *American Economic Review*, 111 (1), 276–323.

——— (2024) "Shared models in networks, organizations, and groups,"Technical report, National Bureau of Economic Research.

Shafer, Glenn (1976) *A mathematical theory of evidence*, 42: Princeton university press.

Shiller, Robert J (2017) "Narrative economics," *American Economic Review*, 107 (4), 967–1004.

Shmaya, Eran and Leeat Yariv (2016) "Experiments on decisions under uncertainty: A theoretical framework," *American Economic Review*, 106 (7), 1775–1801.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman (2011) "How to grow a mind: Statistics, structure, and abstraction," *Science*, 331 (6022), 1279–1285.

Vul, Edward, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum (2014) "One and done? Optimal decisions from very few samples," *Cognitive science*, 38 (4), 599–637.

Vul, Edward and Harold Pashler (2008) "Measuring the crowd within: Probabilistic representations within individuals," *Psychological Science*, 19 (7), 645–647.

Wojtowicz, Zachary (2024) "Model Diversity and Dynamic Belief Formation."

Yang, Jeffrey (2023) "A Criterion of Model Decisiveness."

Supplemental Appendix for "Weighting Competing Models"

A Appendix: Additional Results

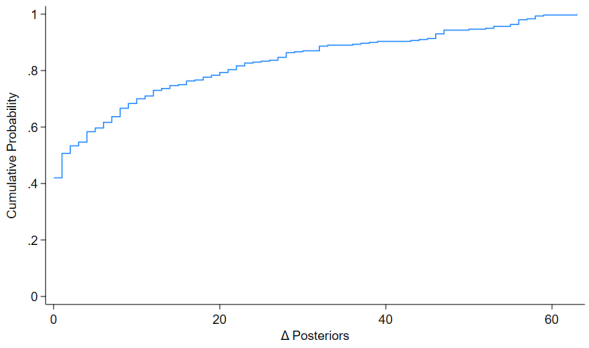
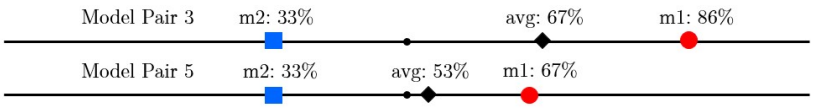
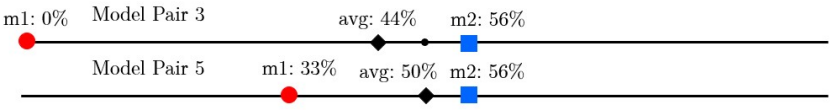


Figure A1: Cumulative Distribution of Within-Task Consistency

Notes. Cumulative distribution of the measure of consistency Δ Posterior, defined as the absolute difference between the reported posteriors given the same signal for the same model pair.



(a) Conditional on Observing a Purple Ball



(b) Conditional on Observing an Orange Ball

Figure A2: Comparison of Posteriors across Model Pairs

Notes. The figures represent the model predictions and average reported guesses for Model Pair 3 and Model Pair 5 and each signal realization. The prior over the state is represented by a small black circle; the model predictions of Model 1 are represented by a red circle, while the ones of Model 2 are represented by a blue square. The average reported posterior for each case is described by a black diamond.

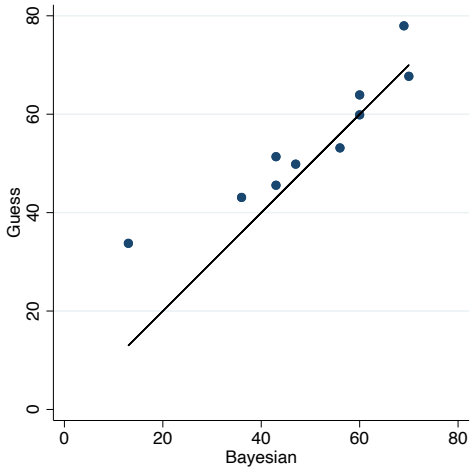


Figure A3: Mean Guesses and Bayesian Predictions

Notes. The figure plots the mean reported guess that bag A is selected and the Bayesian prediction for model pair and signal realization. While the difference between mean belief and the Bayesian prediction is often small, it is statistically significantly different from zero for 8 out of the 10 cases.

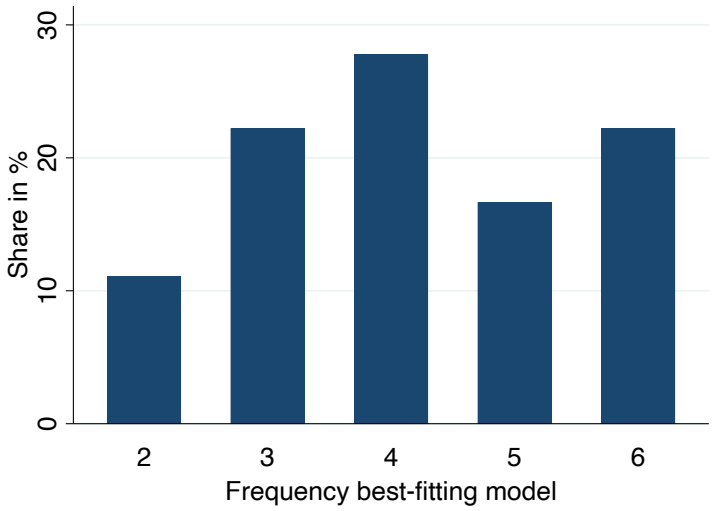


Figure A4: Errors in Model Selection: Robustness

Notes. The figure shows, among the 18 participants who selected a model in 6 out of 7 tasks (see Table 5), how frequently they selected the best-fitting model rather than the worst-fitting one.

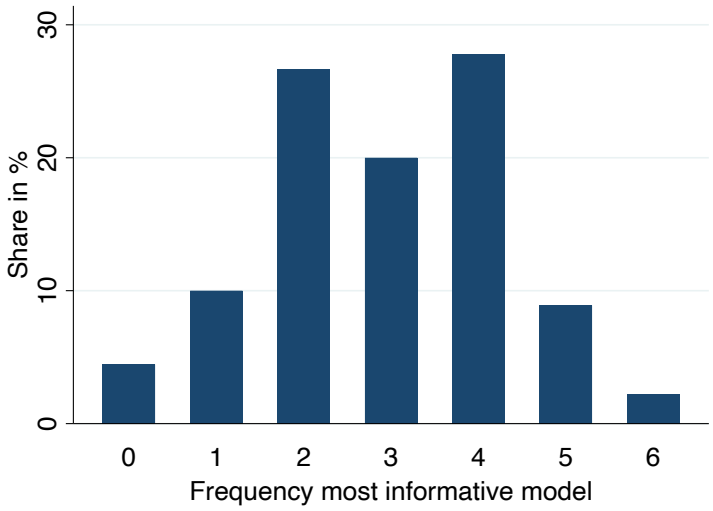


Figure A5: Model Selection based on Informativeness

Notes. The figure shows, among the 90 participants who selected a model in all 7 tasks (see Table 5), how frequently they selected the most informative model.

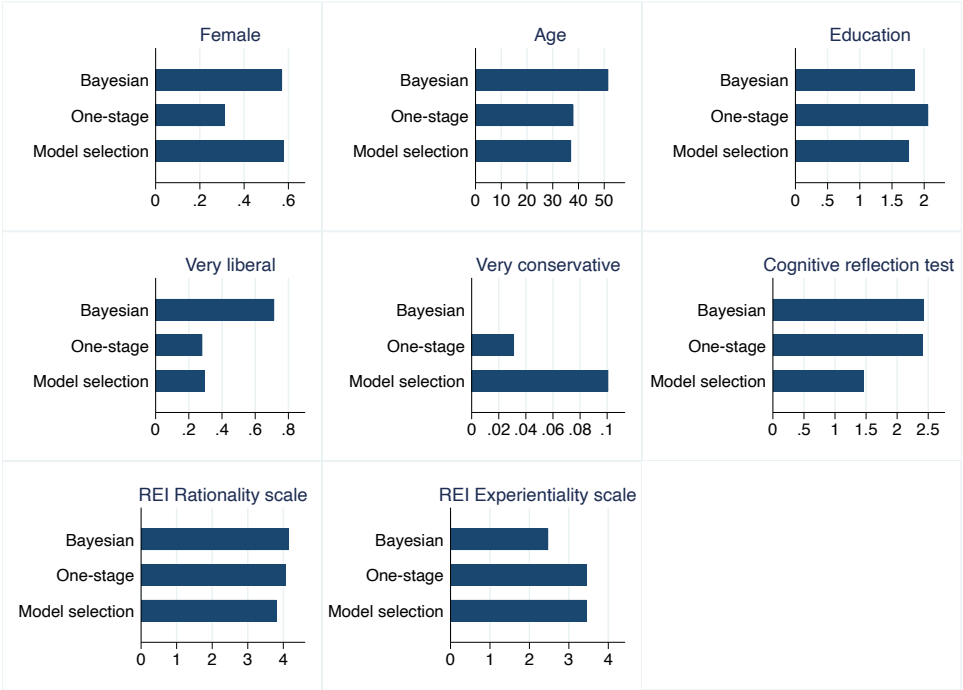


Figure A6: Individual Characteristics by Updating Rules

Notes. The figures show the average individual characteristics of participants who consistently use different updating rules. The sample consists of the 158 participants who consistently used the Bayesian updating, one-stage updating, or model selection in at least 5 out of 7 tasks, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). Due to the small number of observations, we do not include participants who consistently reported the prior. "Female", "Very Liberal" and "Very Conservative" are indicator variables, "Age" is in years (19 and 70), "Education" is from 1 = "High school" to 4 = "PhD or higher", "Cognitive Reflection Test" are scores from 0 to 3, and the Rational Experiental Inventory (REI) rationality and experientiality scales range from 1 to 5.

Type of Guess	% Exact	Within 2 p.p.
Bayesian	3.15	5.90
One-stage	16.51	23.20
Best-fitting	42.86	47.84
Worst-fitting	12.32	13.37
Within	17.56	
Outside	7.60	
Total	100.00	88.21

Table A1: Classification of Guesses: Consistent Guesses

Notes. Consistent participants are classified as reporting a guess within 2 p.p. for the same task as described in Section 4.1. Note that 1.7% of reported guesses are classified both as Bayesian and as one-stage updating, not included in the total.

Dependent Variable	Selects Model 1				
	(1)	(2)	(3)	(4)	(5)
Best-fitting	0.454***		0.445***		0.447***
	(0.038)		(0.040)		(0.040)
Most Informative		-0.165***	-0.032		0.006
		(0.036)	(0.033)		(0.078)
Model Pair 2				0.130***	0.098**
				(0.045)	(0.039)
Model Pair 3				-0.036	0.007
				(0.055)	(0.090)
Model Pair 4				0.037	-0.011
				(0.064)	(0.063)
Model Pair 5				-0.056	0.002
				(0.060)	(0.076)
Constant	0.252***	0.538***	0.268***	0.453***	0.224***
	(0.024)	(0.021)	(0.031)	(0.043)	(0.038)
Observations	956	956	956	956	956
R^2	0.206	0.025	0.207	0.021	0.214

Table A2: Criteria for Model Selection (Within 2 p.p.)

Notes. The table shows coefficient estimates from linear regressions of an indicator of selecting Model 1 on different explanatory variables. The sample consists of the 956 guesses that correspond to model selection allowing for a distance of at most 2 p.p. between the prediction and the reported guess (see Table 2). "Best-fitting" is an indicator for Model 1 having a higher fit than Model 2 given the observed signal. "Most informative" is an indicator for Model 1 being more informative about the state than Model 2 given the observed signal. "Model Pair 2" to "Model Pair 5" are model pair fixed effects that capture any factors that are unconditional on the signal. Standard errors are clustered on the individual level (179 clusters) and are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Type of Vector	Exact		Within 2 p.p.		Within 5 p.p.	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	3.38	[0.43, 6.32]	4.73	[1.27, 8.19]	7.43	[3.16, 11.71]
One-stage	4.73	[1.27, 8.19]	7.43	[3.16, 11.71]	10.81	[5.75, 15.87]
Best-fitting	19.59	[13.12, 26.06]	25.00	[17.94, 32.06]	26.35	[19.17, 33.53]
Worst-fitting	1.35	[0.00, 3.23]	2.03	[0.00, 4.32]	2.03	[0.00, 4.32]
Model 1	4.73	[1.27, 8.19]	6.76	[2.67, 10.85]	6.76	[2.67, 10.85]
Model 2	4.05	[0.84, 7.27]	4.05	[0.84, 7.27]	4.05	[0.84, 7.27]
Prior	2.70	[0.06, 5.35]	2.70	[0.06, 5.35]	2.70	[0.06, 5.35]
Total	40.53		52.70		60.13	

Table A3: Classification of Vectors of Posteriors (Model Pair 2)

Notes. The table reports the shares of vectors of posteriors that have an Euclidean distance from the prediction vector of posteriors or either 0 (Column "Exact"), 2 p.p. (Column "Within 2 p.p.") or 5 p.p. (Column "Within 5 p.p."). We only include the data from Model Pair 2.

Type of Vector	Exact		Within 2 p.p.		Within 5 p.p.	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	1.97	[0.00, 4.21]	1.97	[0.00, 4.21]	3.29	[0.42, 6.16]
One-stage	7.24	[3.07, 11.40]	11.84	[6.65, 17.04]	14.47	[8.82, 20.13]
Best-fitting	19.74	[13.34, 26.14]	21.71	[15.08, 28.34]	23.03	[16.26, 29.80]
Worst-fitting	3.29	[0.42, 6.16]	3.29	[0.42, 6.16]	3.29	[0.42, 6.16]
Model 1	3.95	[0.82, 7.08]	3.95	[0.82, 7.08]	4.61	[1.24, 7.98]
Model 2	5.26	[1.67, 8.85]	7.89	[3.56, 12.23]	7.89	[3.56, 12.23]
Prior	1.32	[0.00, 3.15]	1.32	[0.00, 3.15]	1.32	[0.00, 3.15]
Total	42.77		51.97		57.90	

Table A4: Classification of Vectors of Posteriors (Model Pair 3)

Notes. The table reports the shares of vectors of posteriors that have an Euclidean distance from the prediction vector of posteriors or either 0 (Column "Exact"), 2 p.p. (Column "Within 2 p.p.") or 5 p.p. (Column "Within 5 p.p."). Note that in Column "Within 5 p.p.", 0.33% of reported guesses are classified both as Bayesian and one-stage. We only include the data from Model Pair 3.

	Bayesian	One-stage	Best-fitting	Worst-fitting	Other
Bayesian	2.33	0.00	0.00	0.00	1.33
One-stage	1.33	11.67	0.00	0.00	2.67
Best-fitting	0.67	1.00	28.67	2.67	7.33
Worst-fitting	0.00	0.33	3.67	4.33	1.33
Other	2.67	3.33	3.00	1.00	20.67

Table A5: Classification of Guesses Repeated Model Pair, Same Signal (Within 2 p.p.)

Notes This table uses data from the repeated model pairs (Model Pair 2 and Model Pair 3) for cases where the individual observed the same signal. For each pair of guesses, we classify participants and allow for a distance of at most 2 p.p between the prediction and the reported guess.

Nr. Consistent Observations	Bayesian	One-stage	Model Selection	Prior
0	91.00	72.33	40.33	68.33
1	5.67	13.33	10.33	28.00
2	1.00	3.33	6.00	1.67
3	0.67	2.67	5.00	0.33
4	0.00	1.00	3.33	0.33
5	0.00	1.67	4.67	0.33
6	0.00	3.33	7.67	0.33
7	1.67	2.33	22.67	0.67
Total	100.00	100.00	100.00	100.00

Table A6: Consistency of Updating Rules (Exact)

Notes. The table reports how often different participants use specific updating rules. Since participants complete 7 updating tasks, they can apply each rule between 0 and 7 times (Column "Nr. Consistent Observations"). We require an exact match between the prediction and the reported guess. The columns display the distribution of frequencies for the Bayesian updating, one-stage updating, model selection, and reporting the prior. For example, the column "Bayesian" shows the share of participants who report guesses that correspond to the Bayesian rule in 0, 1, 2, 3, 4, 5, 6, and 7 tasks.

Updating Rule	Mean: Exact	Mean: 2 p.p.	Median: Exact	Median: 2 p.p.
Bayesian	1.67	2.00	1.67	2.67
One-stage	2.33	9.00	8.33	12.00
Model selection	22.67	33.67	38.33	44.00
Prior	0.67	0.67	1.67	1.67
Total	27.33	45.33	50.00	60.33

Table A7: Consistency of Updating Rules: Alternative Approaches

Notes. The table classifies participants according to the updating rules they use in the 7 updating tasks. First, we calculate the distance between the reported guess and each prediction for every observation. For each participant, we then calculate the average and median distance to each prediction. In the column "Mean: Exact," we classify participants if their mean distance to an updating rule is 0. In the column "Mean: 2 p.p.," we allow the average distance of at most 2 p.p.. "Median: Exact" and "Median: 2 p.p." employ a similar approach but use the median distance. Note that "Best- or worst-fitting" corresponds to selecting either the best- or worst-fitting model. The row "Total" gives the share of participants we can classify in these updating rules.

	Bayesian	One-stage	Best-fitting	Worst-fitting	Other	Total
Bayesian	2.67	0.00	0.00	0.00	3.33	6.00
One-stage	1.00	10.00	1.00	0.67	4.33	17.00
Best-fitting	0.67	3.33	19.33	6.67	5.00	35.00
Worst-fitting	0.00	0.67	6.67	4.67	2.33	14.33
Other	0.67	5.67	3.67	2.67	15.00	27.67
Total	5.00	19.67	30.67	14.67	30.00	100.00

Table A8: Classification of Guesses in Task 5 and Task 9 (Within 2 p.p.)

Notes This table uses data from the first updating task (rows) and the last updating tasks before repetitions (columns) with multiple models. For each pair, we classify participants and allow for a distance of at most 2 p.p. between the prediction and the reported guess.

Dependent Variable	Model Selection								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.153**								0.117*
	(0.066)								(0.066)
Age		-0.004							-0.004
		(0.003)							(0.003)
Education			-0.084*						-0.032
			(0.044)						(0.042)
Very Liberal				-0.022					-0.035
				(0.073)					(0.068)
Very Conservative					0.209**				0.096
					(0.082)				(0.086)
Cognitive Reflection Test						-0.133***			-0.115***
						(0.024)			(0.026)
REI Rationality Scale							-0.140***		-0.085
							(0.054)		(0.053)
REI Experientiality Scale								0.045	-0.005
								(0.042)	(0.038)
Observations	181	181	181	181	181	181	181	181	181
R^2	0.030	0.010	0.021	0.001	0.015	0.116	0.031	0.007	0.167

Table A9: Explaining Model Selection: Robustness

Notes. The table shows coefficient estimates from linear regressions of an indicator of model selection on different explanatory variables. The sample consists of the 181 participants who consistently used one updating rule (Bayesian, one-stage, model selection, prior) in at least 4 out of 7 tasks, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). "Female," "Very Liberal," and "Very Conservative" are indicator variables, "Age" is in years (19 and 70), "Education" is from 1 = "High school" to 4 = "PhD or higher," "Cognitive Reflection Test" are scores from 0 to 3, and Rational Experiental Inventory (REI) rationality and experientiality scales are from 1 to 5. Heteroskedasticity-robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent Variable	Model Selection								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.168**								0.144**
	(0.073)								(0.072)
Age		-0.006*							-0.006**
		(0.003)							(0.003)
Education			-0.075						-0.012
			(0.048)						(0.043)
Very Liberal				-0.031					-0.065
				(0.079)					(0.074)
Very Conservative					0.198**				0.098
					(0.089)				(0.097)
Cognitive Reflection Test						-0.131***			-0.109***
						(0.026)			(0.028)
REI Rationality Scale							-0.173***		-0.120*
							(0.061)		(0.062)
REI Experientiality Scale								0.051	-0.001
								(0.047)	(0.043)
Observations	147	147	147	147	147	147	147	147	147
R^2	0.036	0.027	0.017	0.001	0.015	0.110	0.046	0.009	0.194

Table A10: Explaining Model Selection: Robustness

Notes. The table shows coefficient estimates from linear regressions of an indicator of model selection on different explanatory variables. The sample consists of the 147 participants who consistently used one updating rule (Bayesian, one-stage, model selection, prior) in at least 6 out of 7 tasks, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). "Female," "Very Liberal," and "Very Conservative" are indicator variables, "Age" is in years (19 and 70), "Education" is from 1 = "High school" to 4 = "PhD or higher," "Cognitive Reflection Test" are scores from 0 to 3, and Rational Experiental Inventory (REI) rationality and experientiality scales are from 1 to 5. Heteroskedasticity-robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent Variable	Model Selection								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.121								0.088
	(0.080)								(0.077)
Age		-0.009**							-0.008**
		(0.004)							(0.003)
Education			-0.089*						-0.013
			(0.051)						(0.043)
Very Liberal				-0.114					-0.109
				(0.089)					(0.079)
Very Conservative					0.182*				0.089
					(0.097)				(0.104)
Cognitive Reflection Test						-0.148***			-0.114***
						(0.027)			(0.029)
REI Rationality Scale							-0.196***		-0.160**
							(0.065)		(0.063)
REI Experientiality Scale								0.080	0.019
								(0.049)	(0.045)
Observations	121	121	121	121	121	121	121	121	121
R^2	0.019	0.059	0.026	0.015	0.014	0.140	0.063	0.025	0.257

Table A11: Explaining Model Selection: Robustness

Notes. The table shows coefficient estimates from linear regressions of an indicator of model selection on different explanatory variables. The sample consists of the 121 participants who consistently used one updating rule (Bayesian, one-stage, model selection, prior) in 7 out of 7 tasks, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). "Female," "Very Liberal," and "Very Conservative" are indicator variables, "Age" is in years (19 and 70), "Education" is from 1 = "High school" to 4 = "PhD or higher," "Cognitive Reflection Test" are scores from 0 to 3, and Rational Experiental Inventory (REI) rationality and experientiality scales are from 1 to 5. Heteroskedasticity-robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent Variable	Nr. Model Selection								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.795**								0.669*
	(0.345)								(0.352)
Age		-0.024							-0.030*
		(0.015)							(0.015)
Education			-0.564***						-0.372*
			(0.217)						(0.219)
Very Liberal				0.146					0.026
				(0.389)					(0.373)
Very Conservative					0.320				0.406
					(0.672)				(0.645)
Cognitive Reflection Test						-0.516***			-0.462***
						(0.152)			(0.156)
REI Rationality Scale							-0.605**		-0.361
							(0.289)		(0.291)
REI Experientiality Scale								-0.008	-0.179
								(0.210)	(0.200)
Observations	300	300	300	300	300	300	300	300	300
R^2	0.017	0.009	0.021	0.000	0.001	0.035	0.013	0.000	0.080

Table A12: Explaining Model Selection: Robustness

Notes. The table shows coefficient estimates from linear regressions of the number of times a participant selected one model on different explanatory variables. The sample consists of all 300 participants. "Female," "Very Liberal," and "Very Conservative" are indicator variables, "Age" is in years (19 and 70), "Education" is from 1 = "High school" to 4 = "PhD or higher," "Cognitive Reflection Test" are scores from 0 to 3, and Rational Experiental Inventory (REI) rationality and experientiality scales are from 1 to 5. Heteroskedasticity-robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

	Biases in Model Weighting				
	Bayesian	Underinference	Overinference	Wrong direc.	Outside
Bayesian	3.05	0.00	0.00	0.00	0.00
Underinference S.	0.00	6.48	6.86	4.52	0.00
Overinference S.	0.00	10.33	17.38	15.00	2.90
Wrong direc. S.	0.00	0.00	19.00	5.24	9.24

Table A13: Biases in Model Weighting and in Inference about the State

Notes. The table presents the joint distribution of biases in inference about the state of the world (rows) and biases in model weighting (columns).

B Appendix: Heterogeneity in Model Pairs

In this appendix, we explore heterogeneity in the use of updating rules across model pairs. Table 1 presents the model pairs participants encountered in the experiment and Figure B1 illustrates their point predictions. We examine whether characteristics of the models that differ across model pairs systematically influence updating. This analysis is exploratory because: i) there are no existing theories that predict how and which features, beyond fit and informativeness, should affect updating, ii) there is no clear and unified way to categorize differences in model pairs, and iii) the primary purpose of employing diverse model pairs across tasks was to test the robustness of our findings.

We already provide some evidence on the potential importance of the model pair's characteristics in Section 4.2.[41] Table 3 investigates whether participants select models based on characteristics of either models or model pairs that are independent of the signal. Looking at columns (4) and (5), we concluded that such signal-independent characteristics do not play a major role in our setting. Hence, model pair's characteristics that are independent of the signal cannot explain well which model participants select, conditional on model selection. However, they may still influence i) which updating rule participants use more generally, and ii) the likelihood of making mistakes in applying a model selection criterion.

To examine these questions, we replicate the classification of reported guesses presented in the main text for the five different model pairs. Table B1 and Table B2 give the classification for each model pair. Table B1 requires an exact match between the predictions of the updating rules and the guesses, while Table B2 allows for a distance between guesses and predictions of at most 2 percentage points. Additionally, Figure B2 presents the estimated model weights for the different model pairs, and Figures B3 and B4 present the distribution of the weights conditional on the pair and the observed signal.

When considering our main four point predictions, we find the classification of the reported guesses and estimated model weights are strikingly similar across model pairs and consistent with the pooled data presented in the main text. For example, in Table B2, we observe that the share of observations corresponding to one-stage updating and model selection ranges from 15% to 21% and 44% to 47%, respectively. Bayesian updating accounts for approximately 5% of all observations, except for Model Pair 5, where we observe a substantially higher proportion of Bayesian updating; however, this outlier can be explained by the overlap between the predictions of the Bayesian and one-stage rules for this model pair. These findings suggest that model pair characteristics do not play a major role in determining which updating rules individuals use. This aligns with the observation that individuals typically apply updating rules consistently across tasks (see Section 4.4) and that they rarely employ multiple different rules (see Section D.4).

[41] The analysis reported in the main text focuses primarily on the characteristics of model pairs defined conditional on the observed signal, such as fit and informativeness. Our findings indicate that the comparison between the fit levels, rather than informativeness, plays a major role in model selection.

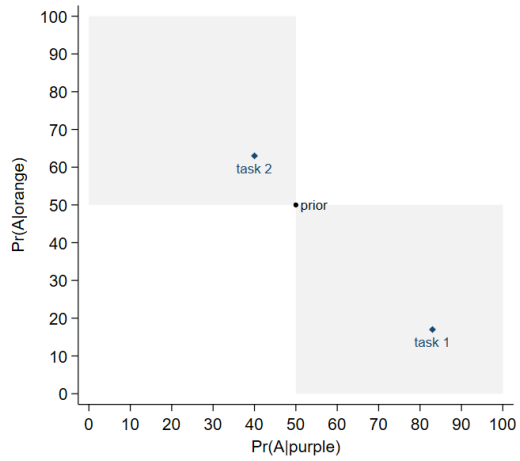
When focusing on participants who engage in model selection, we find that the best-fitting model is selected about 75% of the time. However, Model Pair 4 shows some deviations from this pattern. In this model pair, participants choose the worst-fitting model more frequently than the best-fitting one: only 39% of guesses correspond to the best-fitting model.[42] Model Pair 4 differs substantially from other model pairs as it features a fully uninformative model (equal number of colored balls in each bag) and a fully informative model (each bag contains only balls of a specific color).[43] However, we do not find evidence that participants are systematically drawn to the fully informative nor the fully uninformative model, as shown in Table 3. Therefore, the frequent selection of the worst-fitting model cannot be attributed to the informativeness of the models or any characteristics independent of the signal. While we cannot conclusively explain the pattern in Model Pair 4 and acknowledge that further investigation is needed, the most plausible explanation, also considering our other findings, appears to be based on mistakes in evaluating the fit of models.

Finally, we will explore whether participants are more likely to select the worst-fitting model when the models are close in fit, making it potentially more challenging to identify the best-fitting model. There are some differences in fit levels across model pairs that can be utilized for this purpose, although these differences are relatively small: Model Pair 2 has the smallest fit difference (0.167), followed by Model Pair 5 (0.25), and then Model Pair 1, Model Pair 3, and Model Pair 4 (0.33). However, as Table B2 shows, these differences do not align with the observed variation in the share of selecting the worst-fitting model across tasks.[44]

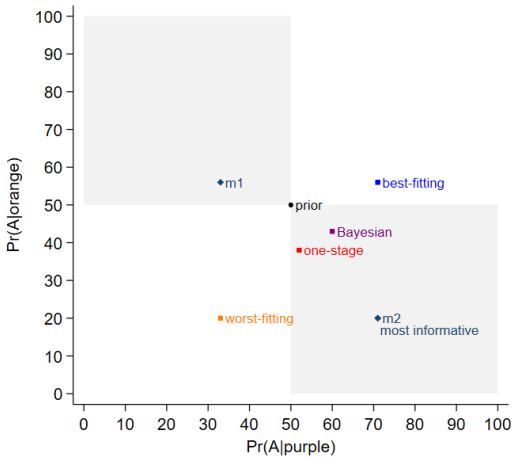
[42]The share of guesses consistent with the best-fitting model does not vary significantly across signals (conditional on model selection: 40% given the orange signal and 38% given the purple signal, p-value = 0.8419; not conditional on model selection: 16% given the orange signal and 20% given the purple signal, p-value = 0.4775).

[43]Moreover, the models are also not conflicting, as defined in Section 2, because the fully uninformative model always implies a posterior equal to the prior. However, the posterior vector resulting from maximum likelihood selection would be Bayes-inconsistent.

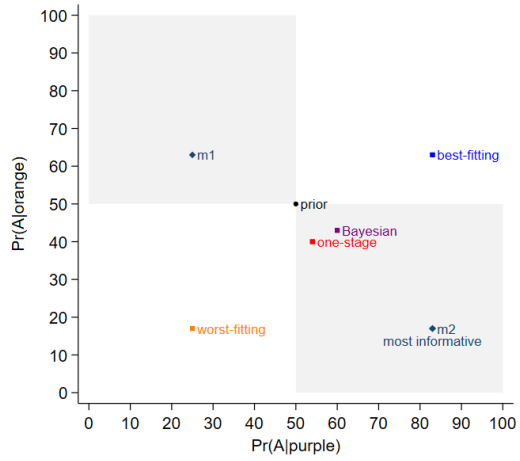
[44]When we regress a dummy for selecting the best-fitting model on the difference in fit levels between proposed models, excluding observations that do not correspond to model selection (within 2 p.p.), we find a statistically significant relationship (N=956, coefficient = -0.675, t = -3.62, $p < 0.001$). However, Model Pair 4 stands out with a disproportionately large share of potential mistakes. When we exclude observations from Model Pair 4, the relationship is no longer statistically significant (N=815, coefficient = -0.043, t = -0.23, p = 0.818). This suggests that other characteristics of Model Pair 4 might explain the frequent potential mistakes. Ultimately, we cannot identify the features that lead to differential updating for this model pair, but we acknowledge that proposing a pair of such distinct models—one fully informative and the other uninformative—appears to be more complex for participants.



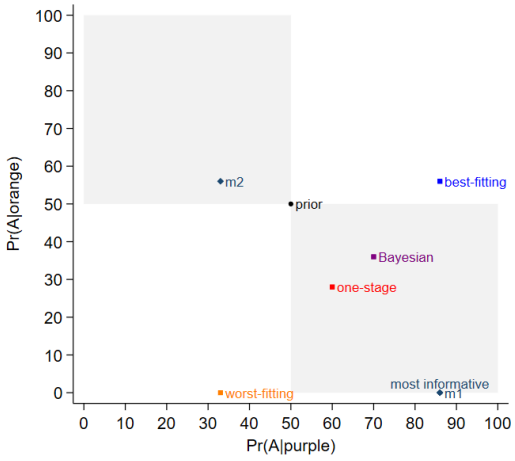
(a) Part 1 and 2



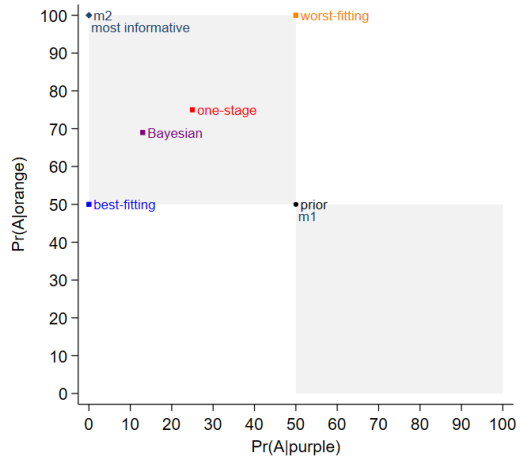
(b) Model Pair 1



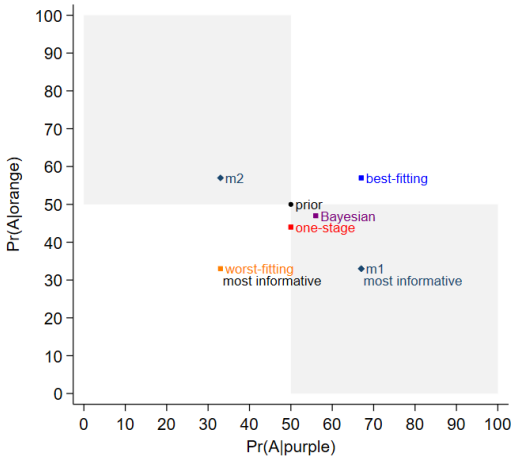
(c) Model Pair 2



(d) Model Pair 3



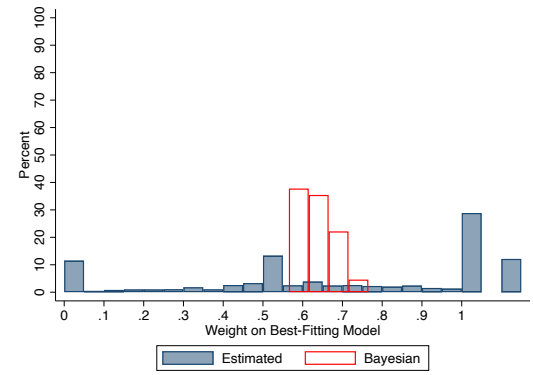
(e) Model Pair 4



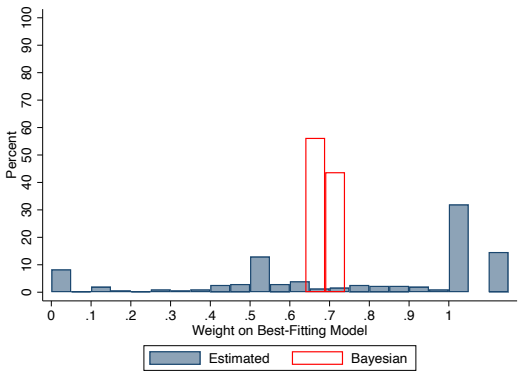
(f) Model Pair 5

Figure B1: Posterior Predictions

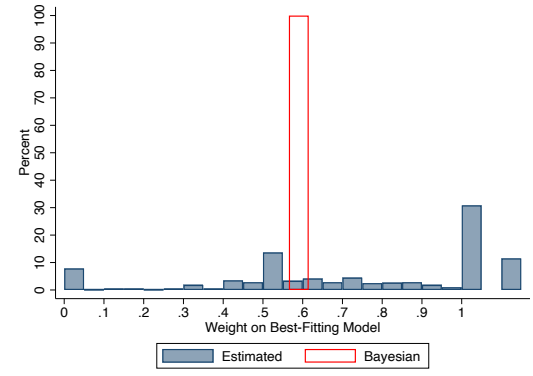
Notes. In all figures, the black circle corresponds to the prior over the state, the blue diamonds to the two model predictions, the purple square to the Bayesian updating for $\Pr(m_1) = 50\%$, the red square to the one-stage updating, the blue square to the prediction by selecting the best-fitting models, and the orange square to the prediction by selecting the worst-fitting models; the gray areas represent the Bayes-consistent vector of posteriors. We also indicate the prediction by selecting the most informative model given each signal.



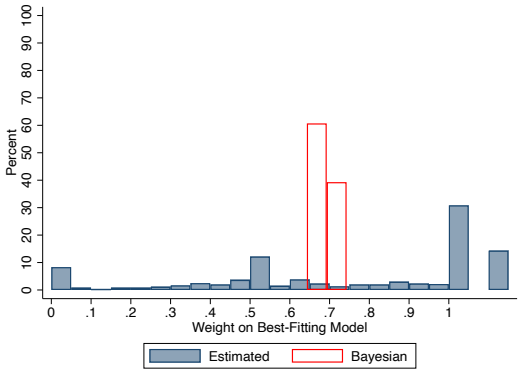
(a) Pooled



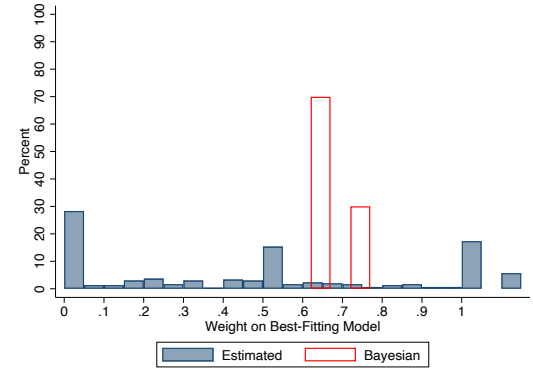
(b) Model Pair 1



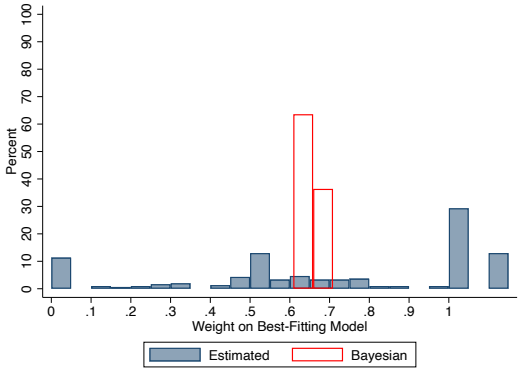
(c) Model Pair 2



(d) Model Pair 3



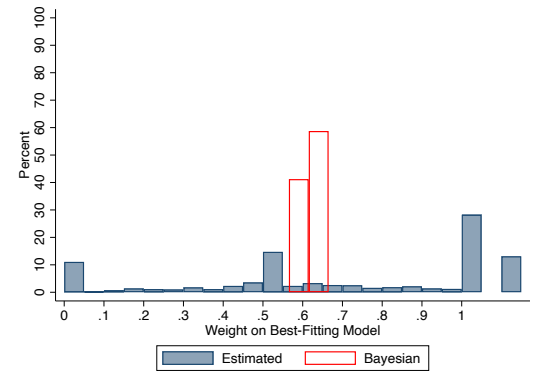
(e) Model Pair 4



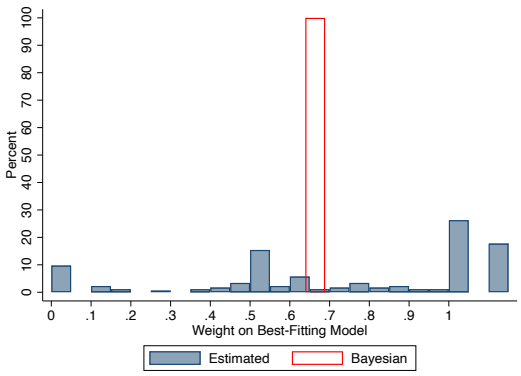
(f) Model Pair 5

Figure B2: Estimated Model Weight

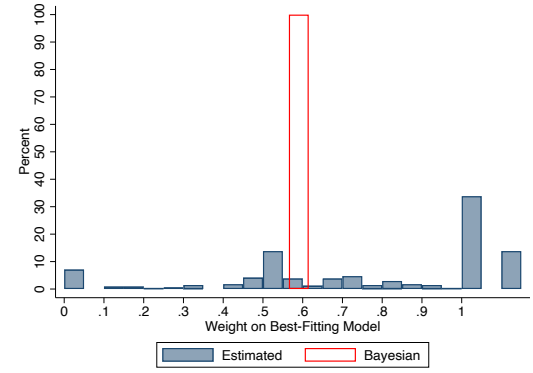
Notes. The figure plots the distribution of weights on the best-fitting model, ρ, for pooled data and for each model pair. We report the implied weights for the reported guesses within the model predictions in blue. The red distribution serves as a benchmark, illustrating how the weights should be distributed if participants follow Bayesian updating in all tasks. The last bar on the right corresponds to the share of reported posteriors outside model predictions for which we cannot recover model weights.



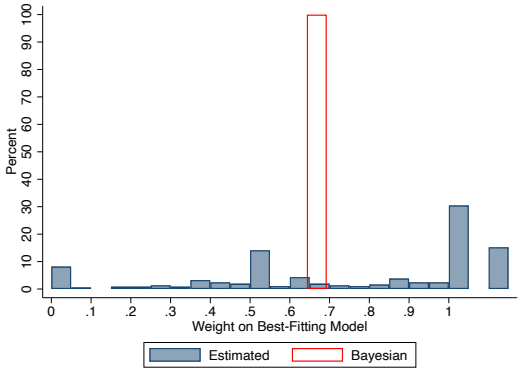
(a) Pooled



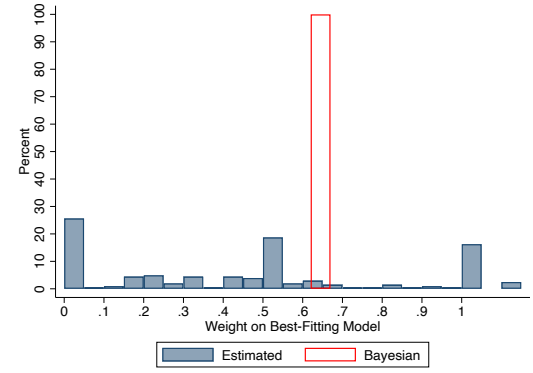
(b) Model Pair 1



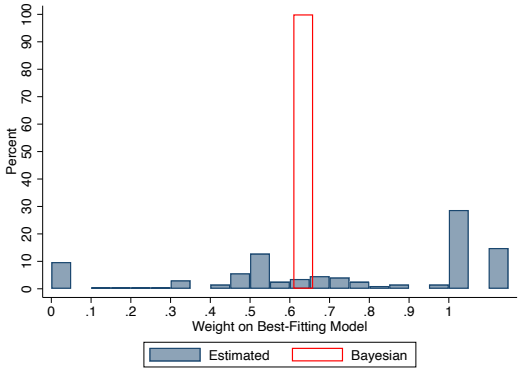
(c) Model Pair 2



(d) Model Pair 3



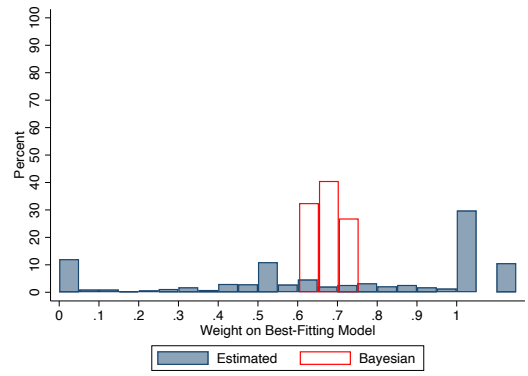
(e) Model Pair 4



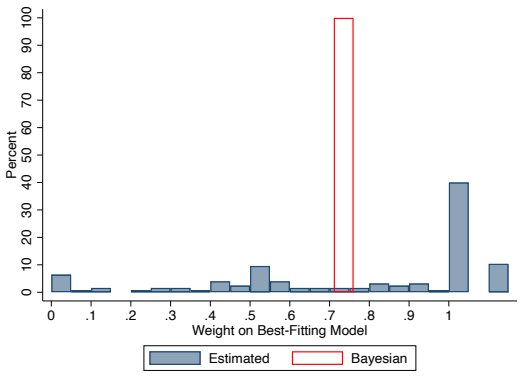
(f) Model Pair 5

Figure B3: Estimated Model Weight (Orange Signal)

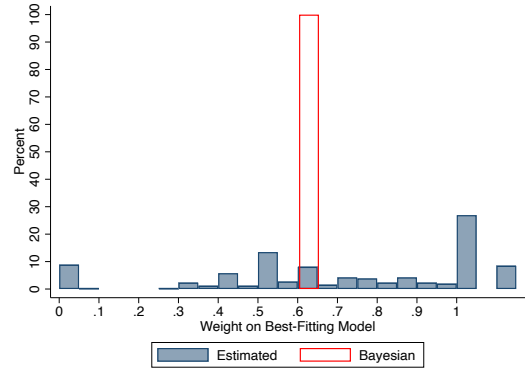
Notes. The figure plots the distribution of weights on the best-fitting model, ρ, for pooled data and for each model pair, conditional on an orange signal. We report the implied weights for the reported guesses within the model predictions in blue. The red distribution serves as a benchmark, illustrating how the weights should be distributed if participants follow Bayesian updating in all tasks. The last bar on the right corresponds to the share of reported posteriors outside model predictions for which we cannot recover model weights.



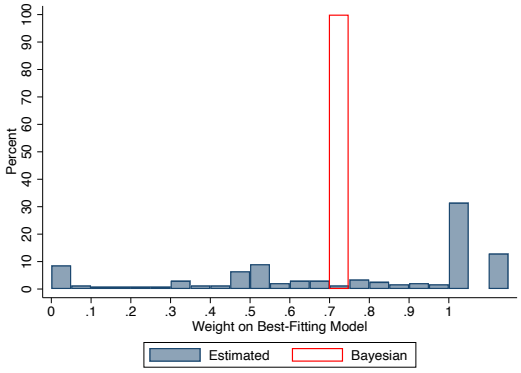
(a) Pooled



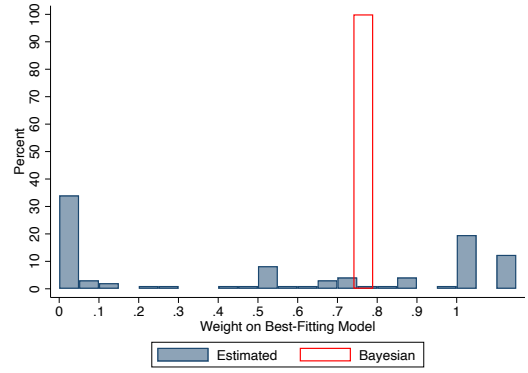
(b) Model Pair 1



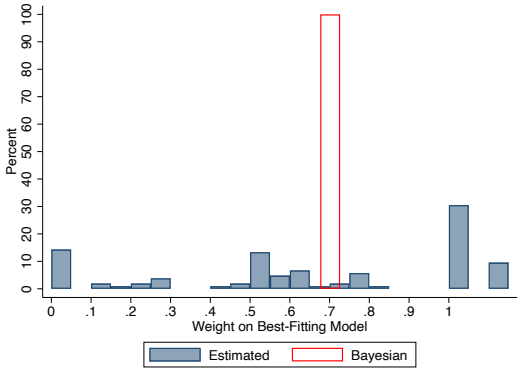
(c) Model Pair 2



(d) Model Pair 3



(e) Model Pair 4



(f) Model Pair 5

Figure B4: Estimated Model Weight (Purple Signal)

Notes. The figure plots the distribution of weights on the best-fitting model, ρ, for pooled data and for each model pair, conditional on a purple signal. We report the implied weights for the reported guesses within the model predictions in blue. The red distribution serves as a benchmark, illustrating how the weights should be distributed if participants follow Bayesian updating in all tasks. The last bar on the right corresponds to the share of reported posteriors outside model predictions for which we cannot recover model weights.

Type of Guess	Model Pair 1	Model Pair 2	Model Pair 3	Model Pair 4	Model Pair 5
Bayesian	2.67	4.00	2.83	2.33	2.67
One-stage	11.00	10.17	9.50	14.33	12.00
Best-fitting	32.00	30.83	30.83	17.33	29.33
Worst-fitting	8.33	7.33	8.00	27.00	10.67
Within	31.33	36.17	34.50	33.33	32.33
Outside	14.67	11.50	14.33	5.67	13.00
Total	100.00	100.00	100.00	100.00	100.00

Table B1: Classification of Guesses by Model Pair: Exact %

Notes. Each column reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, and worst-fitting) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise) for each model pair.

Type of Guess	Model Pair 1	Model Pair 2	Model Pair 3	Model Pair 4	Model Pair 5
Bayesian	6.00	8.83	4.67	5.33	18.67
One-stage	20.33	16.50	15.50	18.33	23.33
Best-fitting	35.67	36.50	36.33	18.33	32.00
Worst-fitting	10.00	9.00	9.33	28.67	11.67
Total	72.00	68.83	65.83	70.66	73.34

Table B2: Classification of Guesses by Model Pair: Within 2 p.p.

Notes Each column reports the shares of guesses that can be classified as one of the point predictions (Bayesian, one-stage, best-fitting, and worst-fitting) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise) for each model pair, allowing a distance of at most 2 p.p. between the prediction and the reported guess. Note that the total does not double-count the overlap of guesses classified both as Bayesian and One-stage: 0% for Model Pair 1, 3, and 4; 2% for Model Pair 2; and 10% for Model Pair 5.

C Appendix: Calibrating a Stochastic Model of Model Selection

This section explores how well our data aligns with a stochastic model of model selection via maximum likelihood. To do so, we estimate the parameters of a simple stochastic model of model selection that includes a heterogeneous population applying different model selection criteria.

Suppose a DM uses one of the criteria for model selection discussed in Section 2: maximum likelihood, informativeness, or dogmatic model selection. The shares of DMs using each criterion are represented by s_{best}, s_{info}, and s_{dogm}, respectively. Also, some DMs could consistently choose the worst-fitting model, with the share s_{worst}. We allow DMs to make errors when applying these criteria, e.g., when comparing the model fit levels. With probability $1-\varepsilon$, the DM selects the model according to their criterion, but with probability ε, they make a mistake and select the other model. We assume ε is constant across participants and updating tasks. This simplifying assumption enables us to perform this exercise and we do not find compelling reasons to adopt a different assumption.

Note that distinguishing between the rate of mistakes ε and the model selection rules is challenging. For example, any posterior belief that corresponds to the prediction of one of the two models can be explained by stochastic versions of model selection based on maximum likelihood or model selection based on informativeness. Furthermore, dogmatic selection imposes minimal structure on individual guesses, allowing it to explain any such posterior without assuming randomness. However, by exploiting the repeated updating tasks, we can identify both s_{best} and ε.

First, we can identify ε from repeated updating tasks when participants observe the same signal twice (see Section 4.1). Participants who choose different models in these two tasks have made one mistake. In our simple model, the proportion of such cases, s_{diff}, corresponds to:

$$s_{\text{diff}} = 2 \cdot \varepsilon \cdot (1-\varepsilon).$$

We estimate s_{diff} from our data, focusing on the 118 observations where a model selection rule was used in both tasks. By using the proportion of participants who select different models in the two updating tasks in the Table A5, we calculate that $s_{diff}= \frac{2.67+3.67}{28.67+2.67+3.67+4.33} = 0.161$. Hence, we estimate ε to be 0.0883.

Next, using this estimated ε value, we can proceed to estimate the shares of DMs following different model selection criteria by using the repeated model pairs when participants observe different signal realizations. Recall from Section 2 that the maximum likelihood criterion requires the selection of the best-fitting model for both signal realizations, which is a different model for the two signals; the same applies to selecting the worst-fitting model. Instead, according to the dogmatic criterion, the participant should select the same model for both signals. Finally, according to the informativeness criterion, partici-

pants should select Model 2 for both signals in Model Pair 2 and Model 1 for both signals in Model Pair 3. Note that dogmatic updating imposes minimal structure on model selection, and for the repeated model pairs we used in our experiment, the dogmatic and informativeness criteria overlap, and thus, we cannot distinguish between them without strong assumptions. Since our main interest is in estimating s_{best}, we combine these criteria under $\mathrm{s}_{\mathrm{same}} = \mathrm{s}_{\mathrm{info}} + \mathrm{s}_{\mathrm{dogm}}$.

We aim to identify the proportion of participants who, in principle, follow a specific criterion but make mistakes, based on the share of participants who report certain vectors of posterior beliefs. We denote the proportions of participants who chose the best-fitting model, the worst-fitting model, and the same model for both signals as $\hat{s}_{\mathrm{best}}, \hat{s}_{\mathrm{worst}}$, and $\hat{s}_{\mathrm{same}}$, respectively. Our simple model posits that these observable shares are as follows:

$$\begin{aligned}
\hat{s}_{\mathrm{best}} &= (1-\varepsilon)^2 \cdot s_{\mathrm{best}} + \varepsilon^2 \cdot s_{\mathrm{worst}} + (1-\varepsilon) \cdot \varepsilon \cdot s_{\mathrm{same}}, \\
\hat{s}_{\mathrm{worst}} &= \varepsilon^2 \cdot s_{\mathrm{best}} + (1-\varepsilon)^2 \cdot s_{\mathrm{worst}} + (1-\varepsilon) \cdot \varepsilon \cdot s_{\mathrm{same}}, \\
\hat{s}_{\mathrm{same}} &= 2 \cdot (1-\varepsilon) \cdot \varepsilon \cdot (1 - s_{\mathrm{same}}) + (\varepsilon^2 + (1-\varepsilon)^2) \cdot s_{\mathrm{same}}.
\end{aligned}$$

Because $\hat{s}_{\mathrm{best}} + \hat{s}_{\mathrm{worst}} + \hat{s}_{\mathrm{same}} = 1$ and $\mathrm{s}_{\mathrm{best}} + \mathrm{s}_{\mathrm{worst}} + \mathrm{s}_{\mathrm{same}} = 1$, we have a system of two equations in two unknowns. We calculate the different shares from our data by focusing on participants who select a model in both updating tasks (see Table 4, Column "Within 2 p.p."). We find that $\hat{s}_{\mathrm{best}} = \frac{23.33}{5.33+6+23.33+2.67} = 0.625$ and $\hat{s}_{\mathrm{same}} = \frac{5.33+6}{5.33+6+23.33+2.67} = 0.3035$. Based on these estimates, we calculate the following shares of participants following the different model selection criteria: $\mathrm{s}_{\mathrm{best}} = 0.7310$, $\mathrm{s}_{\mathrm{worst}} = 0.0588$, and $\mathrm{s}_{\mathrm{same}} = 0.2102$.

Therefore, using this simple calibration exercise, we conclude that model selection can be well described by a stochastic model of model selection via maximum likelihood, with a small minority of participants using other selection rules. While this calibration exercise is based on a very simple model, we believe that it provides a reliable rough estimate of the share of participants who engage in model selection via maximum likelihood, when accounting for stochasticity.

D Appendix: Other Updating Rules

In the main test, we focus on four updating rules: Bayesian updating, one-stage updating, model selection, and reporting the prior over the state. We find that these updating rules can capture most of our data. For individual guesses, Table 2 shows that 53.81% of all reported guesses exactly match these predictions, and this share increases to 71.67% when allowing a distance of at most 2 p.p between the prediction and the reported guess. Moreover, Table 4 illustrates that between 41.67% and 59.01% of the reported vectors of posteriors are consistent with these updating rules. Finally, we also find that participants consistently use these rules: 40.34% of participants use the same updating rule for all 7 guesses, and another 13.66% use the same rule for 5 or 6 out of the 7 tasks (Table 5).

In the following, first, we study whether participants systematically use other updating rules, despite the little margin left by the fact that the rules mentioned above capture most guesses. Then, we also examine whether some participants use multiple updating rules across tasks.

D.1 Over- and Underinference about the Models

In this section, we explore the possibility that some participants report beliefs that correspond to over- and underinference about the models. We discuss in Section 2, how model selection via maximum likelihood and one-stage updating can be interpreted as extreme forms of such biases in inference about the models. Here, we investigate less extreme forms of over- or underinference about the models, both theoretically and empirically.

D.1.1 Theoretical Framework

To more broadly capture over- and underinference about the models, we express the weight placed on the model m as

$$\rho_s^m = \frac{1}{1 + \left(\frac{\Pr(s|m')}{\Pr(s|m)}\right)^{\alpha}}, \tag{D2}$$

where the parameter α captures different type of biased inference: $\alpha = 1$ corresponds to Bayesian updating; $\alpha > 1$ corresponds to overinference with special case $\alpha \to \infty$ as model selection via maximum likelihood; $\alpha \in [0, 1)$ corresponds to underinference with special case $\alpha = 0$ as one-stage updating; and $\alpha < 0$ implies updating in the wrong direction. Finally, selecting the worst-fitting model corresponds to $\alpha \to -\infty$. Note that this general approach to the introduction of biases is related to the method proposed by Grether (1980) to capture biases in updating beliefs about state.

Figure D1 illustrates how vectors of posteriors can be classified as reflecting over- or underinference about the models. A DM exhibiting overinference (underinference) about the models for both signals would report a vector of posteriors in the blue (red) area. The colored lines correspond to the predicted beliefs for different levels of α in Equation

D2, assuming a constant α across signals. The main updating rules that we consider are special cases of different biases in inference about the models: best-fitting for overinference, one-stage for underinference, and worst-fitting for inference in the wrong direction.[45] Note that over- and underinfence about the models imply that the reported posteriors always fall within the range of the model predictions.

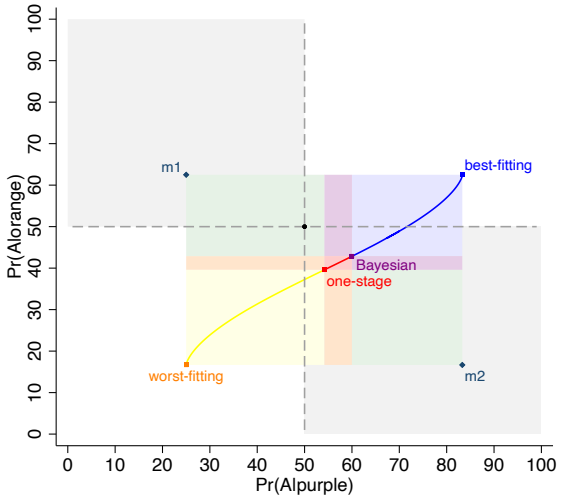


Figure D1: Predictions for Vectors of Posteriors: Biases in Model Weighting

Notes. The figures illustrate two models with the following parameters: $\Pr^{m_1}(p|A) = 1/6$, $\Pr^{m_1}(p|B) = 3/6$, $\Pr^{m_2}(p|A) = 5/6$, and $\Pr^{m_2}(p|B) = 1/6$. In both figures, the black circle corresponds to the prior over the state, the blue diamonds to the two model predictions, the purple square to Bayesian updating with $\Pr(m_1) = 50\%$, the red square to the one-stage updating, the blue square to selecting the best-fitting model, and the orange square to selecting the worst-fitting model; the gray areas represent the Bayes-consistent vectors of posteriors. The colored areas represent the classification for over- and underinference about the models with the following color code: overinference for both signals (blue), underinference for both signals (red), overinference for a signal and underinference for the other signal (purple), wrong direction for both signals (yellow), overinference for a signal and wrong direction for the other signal (green), and underinference for a signal and wrong direction for the other signal (orange). The colored lines capture overinference (blue), underinference (red), and wrong direction (yellow) for a constant degree of α across signals as in Equation D2.

We can generalize this framework to allow for mistakes in evaluating the fit levels of the competing models. If the DM makes a mistake of this type, they sometimes identify the wrong model as the best-fitting one. Therefore, the weight placed on the model m can be expressed more generally as:

$$\rho_s^m = \frac{1}{1 + \left(\frac{\Pr(s|m')}{\Pr(s|m)}\right)^{\lambda\alpha}}, \tag{D3}$$

where the parameter $\lambda \in \{1, -1\}$ determines whether the DM is subject to these mistakes. When $\lambda = 1$, D3 corresponds to deterministic model without mistakes. However, in the stochastic model, with a certain probability, the DM makes an error, resulting in $\lambda = -1$.

[45] Note that model selection based on informativeness and dogmatic model selection are not part of this class of updating rules.

When $\alpha = \infty$, D3 represents the stochastic version of model selection via maximum likelihood, as discussed in Section C.

D.1.2 Results

We begin our analysis by focusing on the deterministic case, where there are no mistakes in identifying the best-fitting model ($\lambda = 1$). Our experimental design allows us to identify α from the individual guesses. When pooling all individual guesses, we find that the most frequent bias is overinference about models, with 43% of guesses falling into this category. Furthermore, we find that 17% of guesses are classified as underinference, while 24% represent inference in the wrong direction. The remaining observations either correspond to Bayesian updating or fall outside the range of the model predictions.

In the main text, we primarily focuse on parameters $\alpha = 1$ (Bayesian updating), $\alpha = 0$ (one-stage updating), and $\alpha \to \infty$ (selecting the best-fitting model) or $\alpha \to -\infty$ (selecting the worst-fitting model). While we have shown that these specific values account for much of the data, here we explore the possibility that some participants consistently report beliefs consistent with other values of α. Figure D2 and Figure D3 present the distribution of α estimated from individual guesses across all tasks and the ones estimated from the vectors of posteriors for the repeated model pairs. These figures do not reveal any frequently used α values other than the ones associated with the main updating rules: $\alpha = 0$ (one-stage updating), $\alpha = 1$ (Bayesian updating), $\alpha = \infty$ (selecting the best-fitting model), and $\alpha = -\infty$ (selecting the worst-fitting model).

Next, we study whether the reported beliefs of some participants can be consistently described by a certain value of α. Therefore, we focus on individuals who do not systematically use any of the main updating rules discussed so far (Bayesian or one-stage updating, model selection, or reporting the prior). Specifically, we consider the 138 participants who used each of these rules in at most 4 tasks, as shown in Table 5. To account for some instability of α, we consider three broad classifications: overinference about the models captured by $\alpha \in (1, \infty)$, underinference about the models captured by $\alpha \in (0, 1)$, and inference about the models in the wrong direction captured by $\alpha \in (-\infty, 0)$. Table D1 presents the distribution of the frequency with which participants use these different updating rules in the seven updating tasks. It indicates that none of these three broad categories is used consistently.

However, the analysis so far assumed that individuals always correctly identify the best-fitting model. Next, we focus on the stochastic case ($\lambda \in \{1, -1\}$). We again consider the 138 participants who use Bayesian updating, one-stage updating, model selection or report the prior in at most 4 tasks. To account for mistakes in identifying the best-fitting model, as well as some instability of α, we consider two broad classifications: overinference about the models captured by $\lambda\alpha \in (-\infty, -1) \cup (1, \infty)$ and underinference about the models captured by $\lambda\alpha \in (-1, 1)$. Table D2 presents the distributions of participants consistent with these categories. When requiring participants to remain

within the same category across all seven guesses, we classify 4.35% of the remaining sample as overinferring about models and 0.72% as underinferring about models. By allowing participants to deviate from the category in one or two guesses, we classify 15.93% (22 participants) and 34.05% (47 participants) of the remaining participants, respectively. This corresponds to 7.33% and 15.66% of the full sample of 300 participants. These findings suggest that some participants consistently use less extreme forms of over- or underinference about models but occasionally make errors in identifying the best-fitting model.

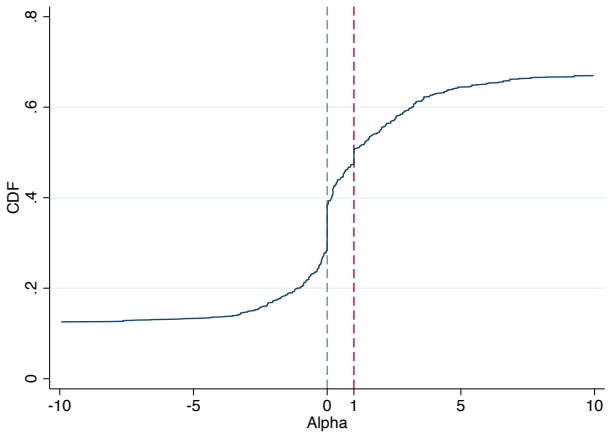
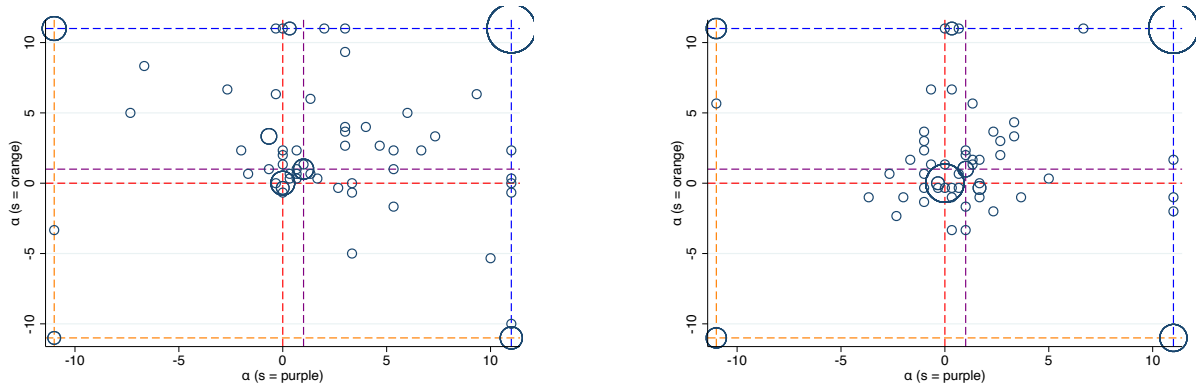


Figure D2: Cumulative Distribution of α

Notes. Cumulative distribution of α across tasks. We exclude observations for which reported guesses fall outside the range of the two model predictions. Note that almost all observations with $\alpha < -10$ and $\alpha^m > 10$ correspond to selecting the worst- or best-fitting model, respectively ($\alpha = -\infty$ or $\alpha = \infty$).



(a) Model Pair 2

(b) Model Pair 3

Figure D3: α in the vectors of Posteriors

Notes. The size of the circles is relative to the number of observations. Data are in steps of 1/3. For this figure, we code selecting the best-fitting or worst-fitting model, corresponding to $\alpha = \infty$ and $\alpha = -\infty$, as 11 and -11.

Nr. Consistent Observations	$\alpha \in (0,1)$	$\alpha \in (1,\infty)$	$\alpha \in (-\infty,0)$
0	49.28	15.22	21.74
1	35.51	30.43	28.99
2	12.32	24.64	27.54
3	2.17	15.94	11.59
4	0.00	6.52	7.97
5	0.00	6.52	2.17
6	0.72	0.72	0.00
7	0.00	0.00	0.00
Total	100.00	100.00	100.00

Table D1: Consistency of Updating Rules: Less Extreme Forms of Over- and Underinference about the Models, Deterministic Case

Notes. The table reports the frequencies of different individuals using specific updating rules. Out of the 7 guesses reported by each participant, we report how many times each rule is used (Column "Nr. Consistent Observations"). The columns display the distribution of frequencies for different levels of α. For example, the column "$\alpha \in (0,1)$" shows the share of participants who report guesses that correspond to any $\alpha \in (0,1)$ in 0, 1, 2, 3, 4, 5, 6, and 7 tasks. The sample consists of the 138 participants who use one of the updating rule (Bayesian, one-stage, model selection, prior) in at most 4 out of 7 tasks, allowing for a distance of at most 2 p.p. between the prediction and the reported guess (see Table 5).

Nr. Consistent Observations	$\lambda\alpha \in (-\infty,-1) \cup (1,\infty)$	$\lambda\alpha \in (-1,1)$
0	1.45	18.12
1	8.70	34.78
2	14.49	16.67
3	26.09	16.67
4	22.46	6.52
5	12.32	5.80
6	10.14	0.72
7	4.35	0.72
Total	100.00	100.00

Table D2: Consistency of Updating Rules: Less Extreme Forms of Over- and Underinference about the Models, Stochastic Case

Notes. The table reports the frequencies of different individuals using specific updating rules. Out of the 7 guesses reported by each participant, we report how many times each rule is used (Column "Nr. Consistent Observations"). The columns display the distribution of frequencies for different levels of $\lambda\alpha$. For example, the column "$\lambda\alpha \in (-\infty,-1) \cup (1,\infty)$" shows the share of participants who report guesses that correspond to any $\lambda\alpha \in (-\infty,-1) \cup (1,\infty)$ in 0, 1, 2, 3, 4, 5, 6, and 7 tasks. The sample consists of the 138 participants who use one of the updating rule (Bayesian, one-stage, model selection, prior) in at most 4 out of 7 tasks, allowing for a distance of at most 2 p.p. between the prediction and the reported guess (see Table 5).

D.2 Signal-independent Model Weights

This section explores the possibility that some participants use model weights that are independent of signals. Examples of such updating rules are dogmatic model selection and one-stage updating. We expand our empirical investigation to consider a broader class of updating rules with this feature, guided by a theoretical framework to capture such updating.

D.2.1 Theoretical Framework

Our goal is to capture the broad class of updating rules in which the model weights are independent of the signal. Such a DM places weight $\rho_s^m = \rho_{s'}^m = c \in [0,1]$ on one model, m, and weight $\rho_s^{m'} = \rho_{s'}^{m'} = 1 - c$ on the other model, m'. This class of rules includes dogmatic model selection with $c \in \{0,1\}$ and one-stage updating with $c = 0.5$.

Building on the graphical representation introduced in Section 2, the green line in Figure D4 illustrates the vectors of posteriors for different values of c. Other updating rules, such as Bayesian updating, predict that ρ_s^m is not signal-independent but depends on the fit levels of the two models conditional on the observed signal s. This also applies to any other updating rules that reveal under- or overinference about the models, as in Section D.1, with the exception of one-stage updating. One-stage updating can be described in terms of both underinference and of signal-independence ($\alpha = 0$ and $c = 0.5$, respectively).

Note that our main analysis is described in terms of the weight placed on the best-fitting model, ρ. In this case, we should expect its value to be in $\{c, 1-c\}$.[46]

D.2.2 Results

Our experimental design allows us to identify the model weights for the reported posteriors that fall within the model predictions. As in the main text, we denote with ρ the weight a participant assigns to the best-fitting model. In the following, we explore the possibility that some participants consistently weight models with $\rho \in \{c, 1-c\}$.

Figure 3 presents the overall distribution of ρ estimated from the individual guesses across all tasks. This figure does not reveal any frequently used ρ values other than the ones associated with the main updating rules: $\rho = 0$ (selecting the worst-fitting model), $\rho = 0.5$ (one-stage updating), and $\rho = 1$ (selecting the best-fitting model).

[46] More precisely, signal-independence implies that $\rho = c$ for one signal and $\rho = 1 - c$ for the other signal, as the best-fitting model always varies across signals. The condition $\rho \in \{c, 1-c\}$ also allows for $\rho = c$ across both signals, which would not reflect signal-independence, but capture over- and underinference about the models as examined in Section D.1, though with a slightly different structural form. This approach would also capture any posteriors lying on the straight line connecting the "best-fitting" and "worst-fitting" points in Figure D4. It is possible to empirically distinguish between these two classes of updating rules only with data on the reported vectors of posteriors. In our empirical approach, we do not differentiate between these two classes of rules, instead considering a broader class of updating rules described by $\rho \in \{c, 1-c\}$. We then present evidence against the relevance of this broader class, providing strong evidence against signal-independent model weighting.

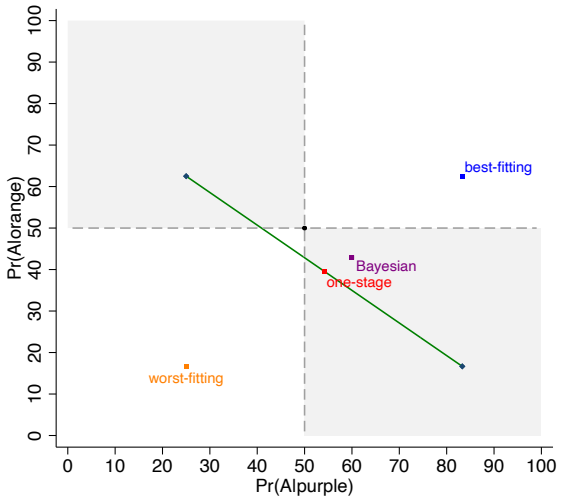


Figure D4: Predictions for Vectors of Posteriors

Notes. The figures illustrate two models with the following parameters: $\Pr^{m_1}(p|A) = 1/6$, $\Pr^{m_1}(p|B) = 3/6$, $\Pr^{m_2}(p|A) = 5/6$, and $\Pr^{m_1}(p|B) = 1/6$. The black circle corresponds to the prior over state, the blue diamonds to the two model predictions, the purple square to Bayesian updating for $\Pr(m_1) = 50\%$, the red square to the one-stage updating, the blue square to selecting the best-fitting model, and the orange square to selecting the worst-fitting models; the gray areas represent the Bayes-consistent vector of posteriors. The green lines capture vectors of posterior resulting from updating rules where model weights are independent of the signals, for all values of c.

Note, however, that signal-independence does not require ρ to be constant; rather, it requires $\rho \in \{c, 1-c\}$. To examine whether there are any frequently used c values, we can analyze the distribution of $|\rho - 0.5|$. A DM who weights models according to $\rho \in \{c, 1-c\}$ would exhibit a constant $|\rho - 0.5|$ across updating tasks. Figure D5 shows the distribution of $|\rho - 0.5|$, revealing that the only frequent values are 0 (one-stage updating) and 0.5 (selection of the best- or worst-fitting model). Thus, we conclude that there are no other frequent c values consistently used by participants.

Finally, we study whether the reported beliefs of some participants can be consistently described by a value of c such that $\rho \in \{c, 1-c\}$. Therefore, we focus on individuals who do not systematically use any of the main updating rules discussed so far (Bayesian or one-stage updating, model selection, prior). Specifically, we consider the 138 participants who used each of these updating rules in at most 4 tasks, as shown in Table 5.

To account for $\rho \in \{c, 1-c\}$ and allow for some instability of ρ, we consider six broad categories: $\rho \in [0, 0.05] \cup (0.95, 1]$, $\rho \in (0.05, 0.15] \cup (0.85, 0.95]$, $\rho \in (0.15, 0.25] \cup (0.75, 0.85]$, $\rho \in (0.25, 0.35] \cup (0.65, 0.75]$, $\rho \in (0.35, 0.45] \cup (0.55, 0.65]$, and $\rho \in (0.45, 0.55]$. Table D3 presents these results, indicating that none of these six broad categories is used consistently. This finding suggests that participants do not consistently use a constant ρ or $\rho \in \{c, 1-c\}$ other than the ones discussed in the main text.

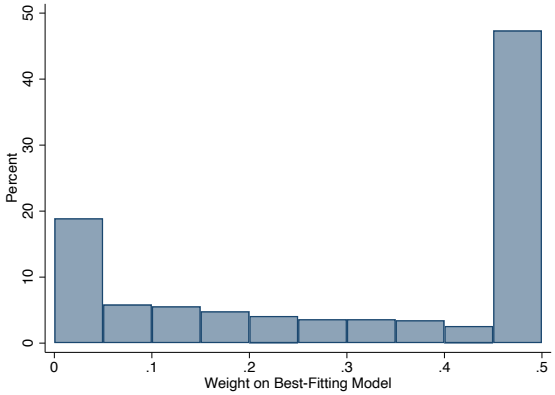


Figure D5: Distribution of $|\rho - 0.5|$

Notes. The figure plots the distribution of $|\rho - 0.5|$

Nr. Consistent Observations	$\rho \in [0, 0.05] \cup (0.95, 1]$	$\rho \in (0.05, 0.15] \cup (0.85, 0.95]$	$\rho \in (0.15, 0.25] \cup (0.75, 0.85]$	$\rho \in (0.25, 0.35] \cup (0.65, 0.75]$	$\rho \in (0.35, 0.45] \cup (0.55, 0.65]$	$\rho \in (0.45, 0.55]$
0	53.62	51.45	47.10	38.41	34.78	46.38
1	18.12	29.71	25.36	32.61	36.96	26.81
2	15.94	14.49	19.57	19.57	15.94	17.39
3	7.97	2.90	6.52	8.70	6.52	6.52
4	4.35	0.72	1.45	0.72	3.62	2.90
5	0.00	0.72	0.00	0.00	2.17	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table D3: Consistency of Updating Rules: Signal-independent Model Weights

Notes The table reports how often different participants use specific updating rules. Since participants complete 7 updating tasks, they can use each rule between 0 and 7 times (Column "Nr. Consistent Observations"). The columns then display the distribution of frequencies for different levels of ρ. For example, the column "$\rho \in (0.05, 0.15] \cup (0.85, 0.95]$" shows the share of participants who report guesses that correspond to any $\rho \in (0.05, 0.15]$ or $\in (0.85, 0.95]$ in 0, 1, 2, 3, 4, 5, 6, and 7 tasks.

D.3 Biases based on the Compound Model

We study belief updating in the presence of competing models by focusing on how individuals weight models to combine their predictions over state. Section 2 presents this approach and the main biases in model weighting that we consider in this project. However, instead of aggregating the predictions of the different models, a DM could first aggregate the competing models into a single model and then use this model to update their beliefs given the observed signal. For a Bayesian DM, these two approaches are equivalent if this compound model is derived by combining the models using the prior over models, $\Pr(m)$. However, as we discuss in this section, biases in updating given such a compound model lead to fundamentally different deviations from the Bayesian benchmark than biases in model weighting as discussed in Section 2. Most importantly, biases based on the compound model reflect the updating biases studied in earlier work (e.g., Benjamin, 2019) and cannot explain the systematic emergence of Bayes-consistencies. For simplicity, we refer to such biases as "biases in updating about the state," distinguishing them from "biases in updating about the models," which are the main focus of our paper.

In this section, we first present a theoretical framework examining biases in updating given the compound model and then study whether this alternative approach can explain the patterns observed in our data. The main goal of this section is to test the descriptive validity of such an approach and to demonstrate that biases in model weighting are empirically distinct from those resulting from updating based on a single model.

D.3.1 Theoretical Framework

In the presence of competing models, a DM responding to a signal could either combine the predictions of the different models, as discussed in Section 2, or aggregate the models into a single model and then use this to update their beliefs. This section focuses on the latter option.

We begin by illustrating that these two approaches are equivalent for a Bayesian DM. In Section 2, we describe the posterior beliefs of a Bayesian DM can be described as:

$$\Pr(A|s) = \rho_s^{m_1} \Pr^{m_1}(A|s) + \rho_s^{m_2} \Pr^{m_2}(A|s), \tag{D4}$$

where the model weight, ρ_s^m, corresponds to $\Pr(m)\Pr(s|m)/\Pr(s)$, starting from prior over the models $\Pr(m)$. Equivalently, the same posterior beliefs could be calculated as:

$$\Pr(A|s) = \frac{\Pr(A)\ \Pr^{\bar{m}}(s|A)}{\Pr^{\bar{m}}(s)} \tag{D5}$$

where $\Pr^{\bar{m}}(s|\omega) = \Pr(m_1)\ \Pr^{m_1}(s|\omega) + \Pr(m_2)\ \Pr^{m_2}(s|\omega)$, for each ω and $\Pr^{\bar{m}}(s) = \Pr(A)\Pr^{\bar{m}}(s|A) + \Pr(B)\Pr^{\bar{m}}(s|B)$. The latter equation allows for an alternative interpretation: the Bayesian DM revises their beliefs via Bayes' rule by using a *compound model*, $\bar{m}$, derived by combining the two models using the prior over the models.

However, a DM may be biased when using the compound model to incorporate the signal into their beliefs. Using Equation D5, it is straightforward to introduce the classical biases in updating when encountering one model. Using the classical approach of Grether (1980), we can captures such biases as:

$$\hat{\Pr}(A|s) = \frac{1}{1 + \left(\frac{\Pr^{\bar{m}}(s|B)}{\Pr^{\bar{m}}(s|A)}\right)^{\alpha_s}}, \tag{D6}$$

where $\alpha_s = 1$ corresponds to Bayesian updating, $\alpha_s \in [0, 1)$ to underinference about the states, $\alpha_s > 1$ to overinference about the states, and $\alpha_s < 0$ to inference in the wrong direction. This classification describes mistakes associated with the misperceived diagnosticity of the observed signal according to the reduced model. For example, underinference represents the case in which the reported posterior is closer to the prior compared to the Bayesian benchmark and thus is "as if" the DM responds to a less diagnostic signal than the observed one. Figure D6 illustrates how these biases could classify different vectors of posterior beliefs, either by allowing α_s values to differ for various signals (colored areas) or by maintaining a constant level of α_s across signals (colored lines).

The comparison between Figure 1b and Figure D6 highlights important differences between biases in model weighting and biases in updating about the state given the compound model. First, even incorporating biases in model weighting, the biased posteriors always fall within the range of the two model predictions. In contrast, biases in updating about the state can result in posteriors that lie outside this range. Second, the predictions of our main updating rules—selecting the best-fitting or worst-fitting model and one-stage updating—cannot be captured by this framework, with the except of Bayesian updating. The predictions from these updating rules not only do not correspond to any specific values of α_s, but they also correspond to inconsistent biases across signal realizations. For example, selecting the best-fitting model would require overinference about the state conditional on one signal and inference in the wrong direction conditional on the other signal. Third, and most importantly, the figure illustrates that biases in updating about the state result in Bayes-consistent vectors of posteriors. This prediction does not depend on individuals using a stable parameter of α_s across signals but rather any combination of $\alpha_s \geq 0$ across signals, that is, when they update in the right direction even if in a biased way. In particular, Bayes-inconsistencies are only predicted when a DM updates in the wrong direction ($\alpha_s < 0$) for only one of the two signals. Such updating does not reflect a consistent updating rule, it should only happen infrequently. As a benchmark, note that a recent study where participants encounter only one model (Aina et al., 2025) finds that the share of Bayes-inconsistent posteriors is quite low (6%). This prediction of biases in updating about the state stands in stark contrast to biases in model weighting, which predict frequent and systematic Bayes-inconsistencies for updating rules such as model selection via maximum likelihood.

Note that the DM may also aggregate the competing models into a single model in a

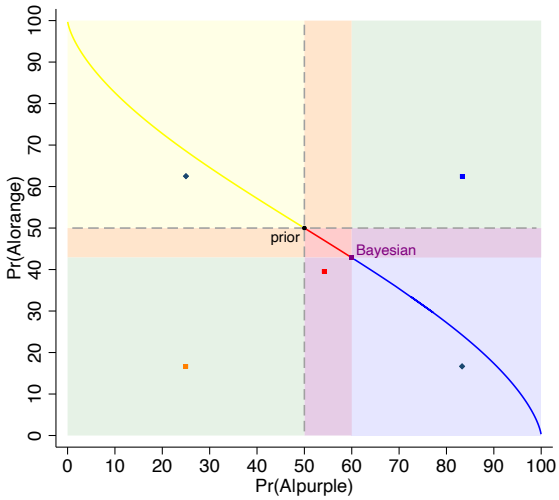


Figure D6: Predictions for Vectors of Posteriors

Notes. The figures illustrate two models with the following parameters: $\Pr^{m_1}(p|A) = 1/6$, $\Pr^{m_1}(p|B) = 3/6$, $\Pr^{m_2}(p|A) = 5/6$, and $\Pr^{m_1}(p|B) = 1/6$. The black circle corresponds to the prior over the state, the blue diamonds to the model predictions, the purple square to Bayesian updating with $\Pr(m_1) = 50\%$, the red square to one-stage updating, the blue square to selecting the best-fitting model, and the orange square to selecting the worst-fitting model; the gray areas represent the Bayes-consistent vector of posteriors. The colored areas represent the classification for over- and underinference, respectively: overinference for both signals (blue), underinference for both signals (red), overinference for a signal and underinference for the other signal (purple), wrong direction for both signals (yellow), overinference for a signal and wrong direction for the other signal (green), and underinference for a signal and wrong direction for the other signal (orange). The colored lines capture overinference (blue), underinference (red), and wrong direction (yellow) for a constant degree of α_s across signals as Equation D6.

biased manner, which could introduce an additional layer of bias that we do not consider in the analysis above. However, crucially, in this case, the DM still only entertains one model and all insights discussed in the previous section still apply. Most importantly, this approach cannot capture Bayesian inconsistencies or updating rules that involve strong over- or underinference about models.

D.3.2 Results

In the main text, we document that a rather small share of participants update their beliefs according to Bayes' rule, which aligns with both biases in model weighting and biases in updating based on a compound model. However, three findings from Section 4 provide evidence in contrast with the idea that our data can be well explained by participants first combine the models into one and then using the latter to update beliefs. Note that these findings not only challenge the framework introduced in Section D.3.1 but also the broader use of a single model to update beliefs. First, we observe that most guesses fall within the range of the two model predictions, which is not predicted by biases in updating about the state. Only 8% of all reported posteriors fall outside the model predictions by more than 2 percentage points. Second, biases in updating about the state can hardly explain why the posteriors of most participants correspond to either model selection or one-stage updating; there are no reasons to expect such belief

patterns to emerge. For example, in the context of the framework of Section D.3.1, such updating would correspond to unstable values of α_s. The frequency of model selection via maximum likelihood is particularly inconsistent with updating based on a single model, as it would require participants to frequently update in the wrong direction, as discussed in Section D.3.1. Third, we find that half of the reported vectors of posteriors are Bayes-inconsistent, at odds with updating based on a single model.

These findings challenge the descriptive validity of any updating rules that rely on aggregating the models into a single model. Hence, they not only question updating rules that aggregate models into a single model using the model prior as a Bayesian DM would do, but also those that aggregate competing models in a biased manner.

We can further explore the potential of more specific updating rules, in which the individual aggregates models using the model prior, but subsequently exhibits systematic biases when updating based on this single model, as discussed in Section D.3.1.[47] To do so, we study whether some participants consistently use some values of α_s other than the ones associated with the updating rules already studied, i.e., $\alpha_s = 0$ (reporting the prior over the state) and $\alpha_s = 1$ (Bayesian updating). Specifically, we consider the 289 participants who used both of these two rules in at most 4 tasks, as shown in Table 5. We then explore whether some of these individuals consistently use another α_s. To account for some instability in α_s, we consider three broad classifications: $\alpha_s \in (1, \infty)$ (overinference about the state), $\alpha_s \in [0, 1)$ (underinference about the states), and $\alpha_s \in (-\infty, 0)$ (inference about the state in the wrong direction). Table D4 reports these results, indicating that a small share of participants use one of these three broad categories of updating rules consistently. This table also illustrates that updating in the wrong direction, $\alpha_s \in (-\infty, 0)$, is frequent. Hence, this class of updating rules fails to adequately capture our data.

This exercise also illustrates a potentially important insight for future research: investigating a setting with competing models while applying the standard approach—where individuals only entertain one model—can lead to incorrect conclusions. Using such an approach, we would conclude that participants' behavior is highly noisy—due to frequent updates in the wrong direction and Bayes-inconsistencies—and that, except for the 11 out of 300 participants who consistently use Bayes' rule or report the prior, updating patterns appear rather unstable. There is no consistent pattern of over- or underinference about the states as almost no participant systematically reports posteriors consistent with a stable parameter α_s.

[47] Our data is not well-suited to study biases in aggregating models into a compound model. Our approach here is to examine deviations from Bayesian updating by fixing the compound model. In principle, we could also do the opposite approach by assuming that participants use Bayes' rule to update their beliefs and then recovering the compound model they entertain—assuming that they indeed aggregate models into a single model. However, this alternative approach is not generally feasible. To apply it, we would also need (i) participants to report vectors of posteriors, and (ii) their vector of posterior beliefs to be Bayes-consistent. We have the reported vectors of posteriors for the repeated model pairs, of which 50% are Bayes-inconsistent. Thus, for half of the sample, there is no compound model that could explain the reported beliefs across signals. Note that this again highlights the difficulty of updating rules that assume participants aggregate models into a single model to explain our data.

Nr. Consistent Observations	$\alpha_s \in [0,1)$	$\alpha_s \in (1,\infty)$	$\alpha_s \in (-\infty,0)$
0	31.14	2.42	19.03
1	38.06	7.96	14.88
2	16.96	19.38	15.22
3	8.30	25.95	23.53
4	4.50	24.57	14.19
5	1.04	12.80	9.69
6	0.00	6.23	3.11
7	0.00	0.69	0.35
Total	100.00	100.00	100.00

Table D4: Consistency of Updating Rules: Over- and Underinference about the State

Notes The table shows how often different individuals use specific updating rules. Since participants complete 7 updating tasks, they can use each rule between 0 and 7 times (Column "Nr. Consistent Observations"). The columns then display the distribution of frequencies for different levels of α_s. For example, Column "$\alpha_s \in [0,1)$" shows the share of participants who report guesses that correspond to any $\alpha_s \in [0,1)$ in 0, 1, 2, 3, 4, 5, 6, and 7 tasks.

D.4 Use of Multiple Rules

In this section, we investigate whether some participants use multiple updating rules among the main rules studied so far. Among the 300 participants, 129 (43%) used one or multiple of the three updating rules—Bayesian updating, one-stage updating and/or model selection—in all 7 tasks (allowing for a distance of at most 2 p.p between the prediction and the reported guess). Additionally, 27 (9.00%) and 20 (6.67%) participants used one or multiple of these rules in 5 and 6 of the 7 updating tasks, respectively.[48] Tables D5,D6 and D7 present the distribution of the combinations of updating rules used by these different groups of participants.

In the following, we refer to a DM's updating pattern as XYZ, where X is the number of times the participant updates according to Bayes' rule (Column "# Bayes" in Table D5), Y denotes the number of times the participant follows the one-stage updating (Column "# One-stage"), and Z gives the number of time the participant selects one model (Column "# One model"). Note that a participant can be classified as using multiple rules in the same task when the predictions of these rules overlap. Consequently, $X + Y + Z$ can exceed 7 (e.g., 170).

Table D5 shows that among participants who use one of the updating rules across all 7 updating tasks, a majority of 69.77% uses model selection in all 7 updating tasks (Row 007). Another 17.83% follow the one-stage updating rule in all 7 tasks (Rows 070 or 170), and 4.66% apply Bayes' rule across all 7 tasks (Rows 700 and 710). Only 7.75% of participants use different rules, with no particular combination of rules being notably frequent: 7 participants (5.43%) use a combination of one-stage updating and model

[48]Furthermore, 25 (8.33%), 37 (12.33%), 29 (9.67%), 22 (7.33%), and 11 (3.67%) participants applied one or multiple of these rules for 4, 3, 2, 1, and 0 of the 7 tasks, respectively.

selection (Rows 016, 025, 043, 061, 134), 1 participant uses a combination of one-stage and Bayes' updating (Row 430), and 2 participants use a combination of all three rules (Rows 224, 261). Tables D6 and D7 also show no frequent or systematic combinations of updating rules among individuals who apply one rule for 6 or 5 out of the 7 tasks. Hence, we conclude that individuals rarely employ multiple rules.

# Bayesian	# One-stage	# Model Selection	# Obs.	Share
0	0	7	90	69.77
0	1	6	3	2.33
0	2	5	1	0.78
0	4	3	1	0.78
0	6	1	1	0.78
0	7	0	14	10.85
1	3	4	1	0.78
1	7	0	9	6.98
2	2	4	1	0.78
2	6	1	1	0.78
4	3	0	1	0.78
7	0	0	1	0.78
7	1	0	5	3.88
			129	100.00

Table D5: Using Different Updating Rules I

Notes The table focuses on the 129 participants who used either Bayes' updating, one-stage updating, or model selection, or a combination of these rules across all 7 updating tasks. We allow for a distance of at most 2 p.p between the prediction and the reported guess. The table then shows the distribution of rule combinations. For example, the first row reports the number (Column "# Obs.") and share (Column "Share") of participants who never used Bayes' rule (Column "# Bayes"), never used the one-stage rule (Column "# One-stage"), and 7 times selected one model (Column "# One model"). Note that Bayesian updating and one-stage updating can produce similar predictions in certain situations, so a single observation might be counted in both the "# Bayes" and "# One-stage" columns.

# Bayesian	# One-stage	# Model Selection	# Obs.	Share
0	0	6	15	55.56
0	1	5	3	11.11
0	2	4	1	3.70
0	3	3	1	3.70
0	6	0	3	11.11
1	0	5	1	3.70
1	6	0	2	7.41
2	0	4	1	3.70
			27	100.00

Table D6: Using Different Updating Rules

Notes The table focuses on the 27 participants who used either Bayes' updating, one-stage updating, or model selection, or a combination of these rules in 6 out of 7 updating tasks. We allow for a distance of at most 2 p.p between the prediction and the reported guess. The table then shows the distribution of rule combinations. For example, the first row reports the number (Column "# Obs.") and share (Column "Share") of participants who never used Bayes' rule (Column "# Bayes"), never used the one-stage rule (Column "# One-stage"), and s selected a model in six tasks (Column "# One model"). Note that Bayes' updating and one-stage updating can produce similar predictions in certain situations, so a single observation might be counted in both the "# Bayes" and "# One-stage" columns.

# Bayesian	# One-stage	# Model Selection	# Obs.	Share
0	0	5	6	30.00
0	1	4	1	5.00
0	2	3	1	5.00
0	5	0	1	5.00
1	0	4	1	5.00
1	1	4	1	5.00
1	2	2	1	5.00
1	4	1	1	5.00
1	5	0	1	5.00
2	1	2	1	5.00
2	2	2	1	5.00
2	3	0	1	5.00
3	3	0	1	5.00
3	3	1	1	5.00
5	1	0	1	5.00
			20	100.00

Table D7: Using Different Updating Rules

Notes The table focuses on the 20 participants who used either Bayes' updating, one-stage updating, or model selection, or a combination of these rules across in 5 of the 7 updating tasks. We allow for a distance of at most 2 p.p between the prediction and the reported guess. The table then shows the distribution of rule combinations. For example, the first row reports the number (Column "# Obs.") and share (Column "Share") of participants who never used Bayes' rule (Column "# Bayes"), never applied the one-stage updating (Column "# One-stage"), and one selected one model in five tasks (Column "# One model"). Note that Bayesian updating and one-stage updating can produce similar predictions in certain situations, so a single observation might be counted in both the "# Bayes" and "# One-stage" columns.

E Appendix: Bias and Response Times

In this section, we discuss deviations from the Bayesian benchmark and the response times associated with different updating rules.

First, we focus on deviations from the Bayesian benchmark, calculated as the absolute distance between the Bayesian prediction and the reported guess. This simple measure describes how different updating rules result in systematic bias in beliefs. Before looking at our data, we calculate the predicted average bias for each updating rule across all tasks. By definition, for a DM consistently applying Bayesian updating, the bias is 0. DMs who consistently use one-stage updating, select the best-fitting model, or select the worst-fitting models are predicted to have an average bias of 6.17 p.p., 15.7 p.p., and 28.24 p.p., respectively. If a DM applies a stochastic version of model selection via maximum likelihood, where mistakes result in selecting the worst-fitting model 8.83% of the time (as estimated in Appendix Section C), the average bias is predicted to be 16.81 p.p..[49]

In Figure E1, we present the empirical distances from the Bayesian benchmark for individuals consistently using Bayesian updating, one-stage updating, or model selection. The figure shows the average absolute distance between guesses and the Bayesian benchmark across all seven updating tasks for the 158 participants who used these rules at least 5 out of 7 times. Consistent with the predicted biases implied by these rules, the average distance is 2.04 p.p. for participants classified as using Bayesian updating, 6.63 p.p. for those using one-stage updating, and 20.52 p.p. for those using model selection.

Figure E2 examines the 90 participants who applied model selection in all seven updating tasks, categorizing them by how often they selected the best-fitting model. The average bias increases monotonically, from 28.38 p.p. for those who never selected the best-fitting model to 16.64 p.p. for those who always did.

This analysis highlights that model selection results in substantially greater deviations from the Bayesian benchmark than one-stage updating, even for participants consistently selecting the best-fitting model (which minimizes the distance to the Bayesian benchmark conditional on using model selection).

These deviations may be acceptable for some participants if they lead to substantial time savings, meaning updating posterior beliefs more quickly. Figure E3 shows the median response times across all seven tasks for individuals consistently using different updating rules. The figure illustrates that both one-stage updating and model selection are associated with substantially shorter response times compared to Bayesian updating: the median response time is 74.62 seconds for participants classified as following Bayesian updating, 19.18 seconds for participants classified as following one-stage updating, and 17.96 seconds for those classified as following model selection. This indicates that DMs face a trade-off between precision and time.

[49] The predicted average distance is 21.66 p.p. using model selection based on informativeness.

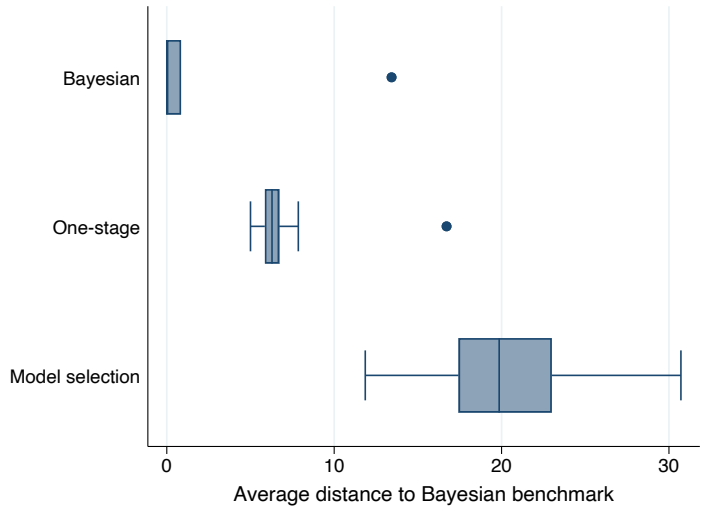


Figure E1: Deviations from the Bayesian Benchmark for Different Updating Rules

Notes. The figure presents a box plot of the average absolute distance between the reported guesses and the Bayesian prediction in all 7 tasks for participants who consistently used one updating rule. The sample consists of the 158 participants that consistently used the Bayesian updating ("Bayesian"), one-stage updating ("One-stage"), or model selection ("Model Selection") in at least 5 out of 7 times, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). Due to the small number of observations, we do not include participants who consistently reported the prior.

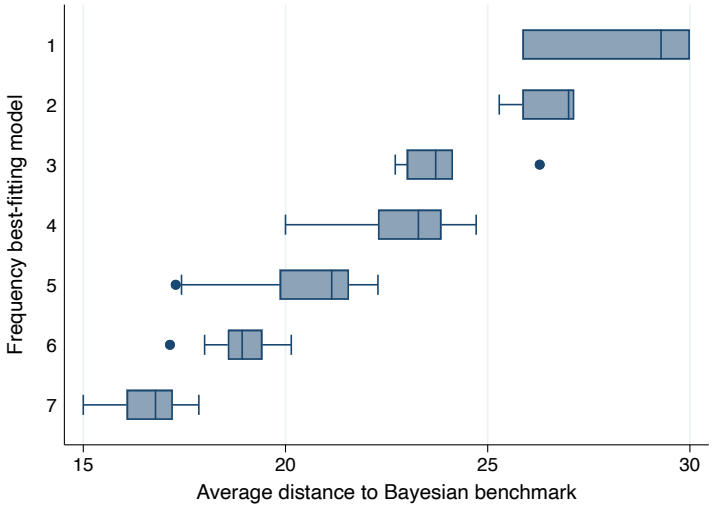


Figure E2: Deviations from the Bayesian Benchmark for Model Selection

Notes. The figure presents a box plot of the average absolute distance between the reported guesses and the Bayesian prediction in all tasks for participants who consistently used model selection. We categorize individuals based on how often they select the best-fitting model (ranging from 0 to 7 times). The sample consists of the 90 participants that consistently select a model in all 7 updating tasks, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5).

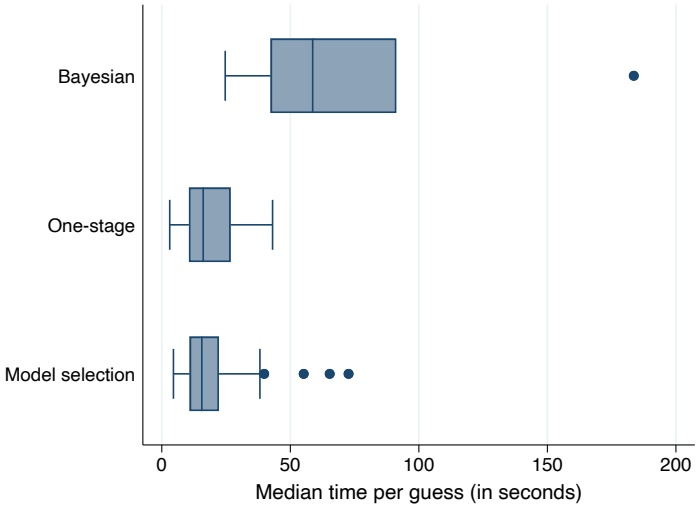


Figure E3: Response Times for Different Updating Rules

Notes. The figure presents a box plot of the median response time in all tasks for participants who consistently used different updating rules. The sample consists of the 158 participants that consistently used the Bayesian ("Bayesian"), one-stage ("One-stage"), or model selection ("Model Selection") updating rules in at least 5 out of 7 times, allowing for a distance of at most 2 p.p between the prediction and the reported guess (see Table 5). Due to the small number of observations, we do not include participants who consistently reported the prior.

F Appendix: Additional Data Collection

This section includes additional information on our additional studies. First, we describe the details of the experimental designs and data collections. We then present additional results and figures. These sections complement Section 5 of the main paper.

F.1 Experimental Design

While Section 5 is organized by treatment, here we proceed by study. For reference, the second study includes the *No-Prior* and *Click* treatments, while the third study includes the *Non-Conflicting* and *Three-Models* treatments.

F.1.1 Second Study

To test the generalizability of our results in other settings, the second study closely replicates the design presented in Section 3, including instructions, with a few modifications that we discuss in this section.

First, we shorten the experiment by reducing the number of updating tasks in Part 3 from seven to two, using Model Pair 2 and Model Pair 3. These pairs are well-suited for identifying updating rules, as the predictions of different rules are particularly distinct (Appendix Figure B1). This is also why we used them as the repeated model pairs in the main study. Participants encounter these two model pairs in random order.

Second, we change how we present the selection of the implemented model to participants. In our main study, we explained to participants that a die roll would randomly determine which model was implemented in that task: if the number 1, 2, or 3 is rolled, Model 1 was selected, while otherwise Model 2 was selected. To avoid restricting possible model priors to a small set in the *No-Prior* condition, we replace the die roll with a card draw in all treatments. The card is drawn from a deck of 100 red and blue cards, where red cards correspond to Model 1 (a red robot) and blue cards correspond to Model 2 (a blue robot). In the *Baseline* and *Click* conditions, participants know that 50 cards are red and 50 cards are blue, implementing the same objective prior over models as in our main study. Instead, in the *No-Prior* condition, participants do not know how many cards are blue or red. A control question verifies the comprehension of this aspect.

Third, since this study required a shorter completion time and half the tasks of our main study, we adjusted the completion fee to 5.7 USD and the bonus payment to 1 USD.

Fourth, we shorten Part 4, where participants fill out a short survey that only includes a modified version of the Cognitive Reflection Test (Frederick, 2005) and demographics.

The only difference between the *No-Prior* and *Baseline* conditions lies in whether the color composition of the cards is known to participants. The *Click* condition differs from *Baseline* in that model predictions are not automatically displayed to participants. Instead, participants can access each prediction by clicking a corresponding button five

times in a row. This feature is introduced in Part 2 of the study, along with the model predictions. In Part 2, participants must press the "G" button to reveal the prediction of the model (represented by a green robot). In Part 3, they must press the "R" button for revealing the prediction of the first model (red robot) and the "B" button for the second model (blue robot). Participants cannot reveal both predictions simultaneously.

Logistics The experiment was pre-registered on AsPredicted (`https://aspredicted.org/wqtd-68zm.pdf`) and conducted on Prolific in February 2025, restricting the participant pool to US residents, aged 18-70, with approval rates of at least 95%. The study was completed through a link to a Qualtrics survey, including instructions and control questions for each part (Appendix G). A total of 592 participants completed the study successfully. Of these participants, 34% and 37% were randomly allocated to *Baseline* and *No-Prior* conditions, respectively. This difference in treatment allocation is purely random: participants can only be excluded for failing attention checks, which take place at the beginning of the study when all treatments are identical. The average payment was 6.5 USD, and the average duration was approximately 30 minutes. In our final sample, 49% are female, 24% have low schooling ('High school' or lower educational level), and the median age is 37.

F.1.2 Third Study

The third study closely replicates the design of the second study, with three differences.

First, we increase the number of updating tasks in Part 3 from two to three. Similar to the second study, we build on Model Pair 2 and Model Pair 3. The two updating tasks are completed in a random order, and then they encounter the same models they encountered in the first task a second time. This design feature allows us to recover the vectors of posterior beliefs for participants who randomly observed different signals in the repeated tasks. As a result, we can study the presence of Bayesian inconsistencies, which is particularly interesting in the context of non-conflicting models.

Second, we adjust how we present the selection of the implemented model to participants. While the second study uses a deck of 100 cards to represent the prior distribution over the models, we reduce the deck to 60 cards in the third study. This modification ensures that in the *Three-model* condition, each of the three models can be drawn with equal probability of 1/3.

Third, we incorporate multiple safeguards to address potential submissions by AI agents, a concern that had grown in Prolific at the time the third study was conducted (but not at the time of the main or second study). These safeguards include captchas—excluding participants with scores below 0.5 as pre-registered—as well as a short audio transcription task and a video-based attention check requiring respondents to correctly enter four numbers displayed in a video, following Çelebi et al. (2025).

Moreover, the treatment conditions in the third study require modifying the original

model pairs relative to those in the *Baseline*. We outline in the main text how these models, presented in Appendix Table F1, are chosen.

Logistics The experiment was pre-registered on AsPredicted (`https://aspredicted.org/78jp45.pdf`) and conducted on Prolific in December 2025, restricting the participant pool to US residents, aged 18-70, with approval rates of at least 95%. The study was completed through a link to a Qualtrics survey, including instructions and control questions for each part (Appendix G). A total of 586 participants completed the study successfully and had captcha scores above 0.5. Of these participants, 35% and 32% were randomly allocated to *Baseline* and *Non-Conflicting* conditions, respectively. This difference in treatment allocation is purely random: participants can only be excluded for failing attention checks, which take place at the beginning of the study when all treatments are identical. The average payment was 6.5 USD, and the average duration was approximately 30 minutes. In our final sample, 50% are female, 29% have low schooling ('High school' or lower educational level), and the median age is 41.

Part	Pair	$\Pr^{m_1}(p\|A)$	$\Pr^{m_1}(p\|B)$	$\Pr^{m_2}(p\|A)$	$\Pr^{m_2}(p\|B)$	$\Pr^{m_3}(p\|A)$	$\Pr^{m_3}(p\|B)$
				Baseline			
3	2	1/6	3/6	5/6	1/6		
3	3	6/6	1/6	1/6	2/6		
				Non-Conflicting			
3	2'	1/6	0/6	5/6	4/6		
3	3'	5/6	3/6	2/6	0/6		
				Three-Models			
3	2"	1/6	3/6	5/6	1/6	3/6	2/6
3	3"	6/6	1/6	1/6	2/6	4/6	1/6

Table F1: Models Used in the Third Study

Notes. This table describes all models used in Part 3 of the third experiment. In the *Baseline* and *Non-Conflicting* conditions, participants face two models (m_1 and m_2). In the *Three-Models* condition, they encounter three models (m_1, m_2, and m_3). In each updating task, there are two bags, representing the state of the world, $\omega \in \{A, B\}$, with each bag containing six balls, either purple (p) or orange (o). A model m sets the share of purple balls in each bag, $\Pr^m(p|A)$ and $\Pr^m(p|B)$, as shown in the corresponding columns.

F.2 Additional Results

Type of Guess	Exact		Within 2 p.p.	
	%	95%-CI	%	95%-CI
Bayesian	3.17	[1.43, 4.90]	5.50	[3.34, 7.66]
One-stage	9.17	[6.25, 12.09]	15.33	[11.68, 18.99]
Best-fitting	32.67	[28.02, 37.31]	38.33	[33.60, 43.07]
Worst-fitting	7.50	[5.17, 9.83]	9.67	[7.11, 12.23]
Within	33.00	[28.46, 37.54]		
Outside	14.50	[11.39, 17.61]		
Total	100.00		68.83	

Table F2: Classification of Guesses by Treatment (Main Study)

Notes. Column "Exact" reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, and worst-fitting) or in one of the two residual categories ("Within" if the guess is within the model predictions or "Outside" otherwise). Column "Within 2 p.p." reports the shares of guesses that fall within 2 p.p. around each point prediction. We pool the data from both Model Pair 2 and Model Pair 3 before the repetitions from the main study (N=600).

F.2.1 Second Study

Type of Guess	Baseline		No-Prior		Click	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	3.98	[1.25, 6.71]	7.76	[4.19, 11.33]	5.81	[2.28, 9.35]
One-stage	13.93	[9.10, 18.76]	16.44	[11.49, 21.39]	17.44	[11.71, 23.17]
Best-fitting	41.29	[34.43, 48.16]	43.38	[36.76, 49.99]	40.70	[33.28, 48.11]
Worst-fitting	13.93	[9.10, 18.76]	10.96	[6.79, 15.13]	11.63	[6.79, 16.47]
Total	73.13		78.54		75.58	

Table F3: Classification of Guesses by Treatment, Model Pair 2 (Second Study)

Notes. The table reports the shares of guesses that fall within 2 p.p. around each point prediction (Bayesian, one-stage, best-fitting, and worst-fitting) for each treatment condition (Baseline, No-Prior, Click). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We only use the data from Model Pair 2.

Type of Guess	Baseline		No-Prior		Click	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	3.48	[0.93, 6.04]	1.37	[0.00, 2.92]	4.07	[1.09, 7.05]
One-stage	10.45	[6.18, 14.71]	10.50	[6.41, 14.59]	14.53	[9.21, 19.86]
Best-fitting	46.77	[39.81, 53.72]	43.84	[37.21, 50.46]	48.26	[40.71, 55.80]
Worst-fitting	8.96	[4.97, 12.94]	9.13	[5.29, 12.98]	11.05	[6.31, 15.78]
Total	69.66		64.84		77.91	

Table F4: Classification of Guesses by Treatment, Model Pair 3 (Second Study)

Notes. The table reports the shares of guesses that fall within 2 p.p. around each point prediction (Bayesian, one-stage, best-fitting, and worst-fitting) for each treatment condition (Baseline, No-Prior, Click). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We only use the data from Model Pair 3.

Type of Guess	Baseline vs. No-Prior				Baseline vs. Click			
	Coeff.	t	p	95%-CI	Coeff.	t	p	95%-CI
Bayesian	0.81	0.57	0.57	[-1.99, 3.60]	1.30	0.85	0.40	[-1.71, 4.31]
One-stage	1.02	0.35	0.73	[-4.67, 6.71]	3.80	1.19	0.24	[-2.49, 10.09]
Best-fitting	-0.53	-0.13	0.90	[-8.74, 7.68]	0.31	0.07	0.95	[-8.64, 9.26]
Worst-fitting	-1.24	-0.53	0.59	[-5.80, 3.32]	-0.08	-0.03	0.98	[-5.21, 5.06]
Model Selection	-1.77	-0.40	0.69	[-10.49, 6.94]	0.24	0.05	0.96	[-9.31, 9.79]
Best-fitting, cond.	1.69	0.43	0.67	[-6.08, 9.47]	0.13	0.03	0.98	[-8.52, 8.77]
Identified	-0.69	-0.18	0.86	[-8.17, 6.78]	4.96	1.26	0.21	[-2.78, 12.71]

Table F5: Treatment Comparisons (Second Study)

Notes. This table presents treatment effects by regressing indicator variables, capturing whether participants are classified under different updating rules (allowing for a 2 p.p.; see Table 7), on treatment dummies. The analysis pools data from both updating tasks, clusters standard errors at the individual level, and includes model pair x signal fixed effects. The column "Baseline vs. No-Prior" compares the *Baseline* and *No-Prior* conditions (N=840), while the column "Baseline vs. Click" compares the *Baseline* and *Click* conditions (N=746). For the row "Model Selection," the outcome variable is defined as 1 if the participant selects either the best- or worst-fitting model and 0 otherwise. For the row "Best-fitting, cond.," we restrict the sample to observations where participants engage in model selection (N=458 for *No-Prior* and N=415 for *Click*). For the row "Identified," the outcome variable is defined as 1 if the observation is classified as Bayesian, One-stage, Best-fitting, or Worst-fitting, and 0 otherwise.

Type of Guess	Baseline vs. No-Prior				Baseline vs. Click			
	Coeff.	t	p	95%-CI	Coeff.	t	p	95%-CI
Bayesian	0.17	0.23	0.82	[-1.24, 1.57]	1.29	1.46	0.14	[-0.45, 3.03]
One-stage	0.01	0.01	1.00	[-4.95, 4.97]	3.16	1.11	0.27	[-2.44, 8.76]
Best-fitting	0.88	0.22	0.83	[-7.05, 8.80]	2.66	0.60	0.55	[-6.03, 11.35]
Worst-fitting	-2.76	-1.29	0.20	[-6.97, 1.45]	-0.80	-0.32	0.75	[-5.76, 4.16]
Model Selection	-1.88	-0.43	0.67	[-10.53, 6.76]	1.86	0.39	0.70	[-7.59, 11.32]
Best-fitting, cond.	5.35	1.25	0.21	[-3.05, 13.74]	2.34	0.48	0.63	[-7.20, 11.87]
Identified	-1.71	-0.40	0.69	[-10.12, 6.71]	6.32	1.40	0.16	[-2.56, 15.20]

Table F6: Treatment Comparisons, Exact (Second Study)

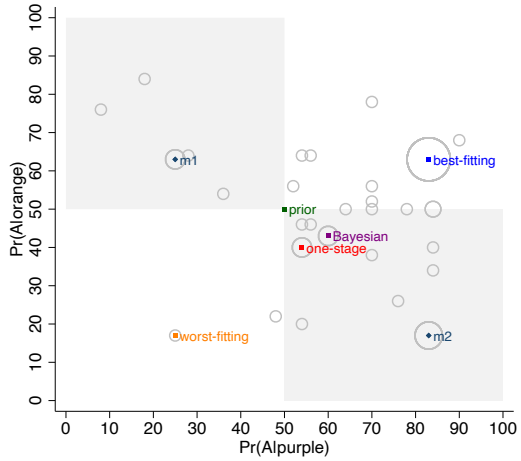
Notes. This table presents treatment effects by regressing indicator variables, capturing whether participants are classified under different updating rules (requiring an exact match; see Table 7), on treatment dummies. The analysis pools data from both updating tasks, clusters standard errors at the individual level, and includes model pair x signal fixed effects. The column "Baseline vs. No-Prior" compares the *Baseline* and *No-Prior* conditions (N=840), while the column "Baseline vs. Click" compares the *Baseline* and *Click* conditions (N=746). For the row "Model Selection," the outcome variable is defined as 1 if the participant selects either the best- or worst-fitting model and 0 otherwise. For the row "Best-fitting, cond.," we restrict the sample to observations where participants engage in model selection (N=458 for *No-Prior* and N=415 for *Click*). For the row "Identified," the outcome variable is defined as 1 if the observation is classified as Bayesian, One-stage, Best-fitting or Worst-fitting, and 0 otherwise.

F.2.2 Third Study

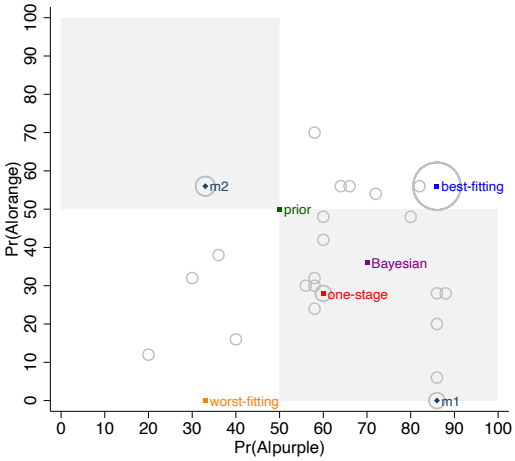
Type of Guess	Baseline		Non-Conflicting		Three-Models	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	4.41	[2.10, 6.73]	23.76	[20.11, 27.41]	16.32	[13.13, 19.52]
One-stage	15.85	[11.43, 20.26]	25.00	[19.58, 30.42]	25.26	[20.43, 30.08]
Best-fitting	48.86	[43.05, 54.66]	46.28	[40.40, 52.15]	32.47	[26.90, 38.05]
Worst-fitting	8.17	[5.51, 10.83]	9.40	[6.59, 12.20]	5.15	[2.77, 7.54]
Middle model					24.40	[20.07, 28.72]
Total	76.80		84.93		72.16	

Table F7: Classification of Guesses by Treatment, Within 2 p.p. (Third Study)

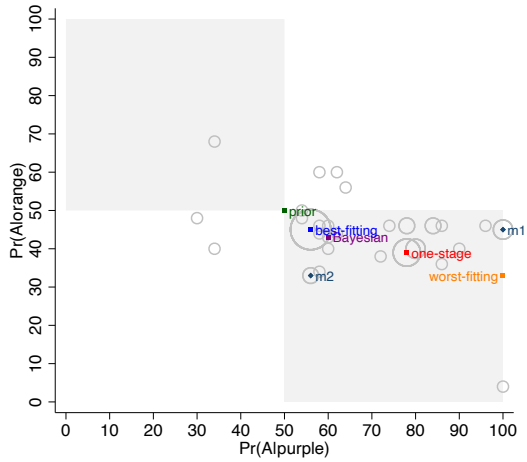
Notes. The table reports the shares of guesses that fall within 2 p.p. around each point prediction for each treatment (*Baseline*, *Non-Conflicting*, *Three-Models*). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We pool the data from all updating tasks. Note that, in *Baseline*, 0.49% of reported guesses are classified both as Bayesian and one-stage, in *Non-Conflicting*, 15.78% of reported guesses are classified both as Bayesian and best-fitting, 2.48% as Bayesian and one-stage, and 1.24% as Bayesian and worst-fitting, in *Three-Models*, 5.15% of reported guesses are classified both as Bayesian and middle model, 9.45% as Bayesian, one-stage and middle model, and 7.39% as both one-stage and middle model. These duplicate observations are not included in the total.



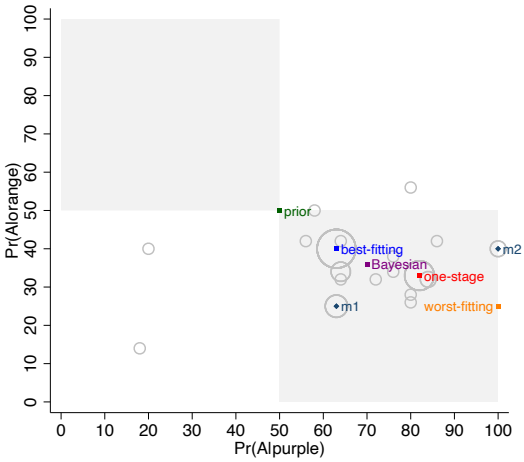
(a) Model Pair 2



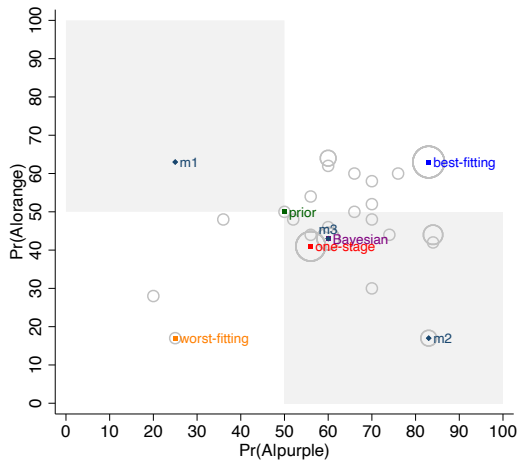
(b) Model Pair 3



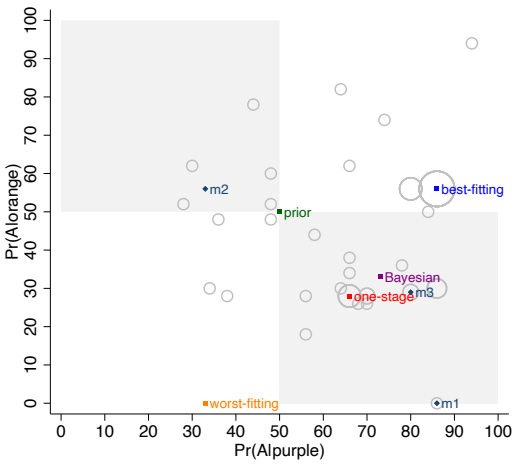
(c) Model Pair 2'



(d) Model Pair 3'



(e) Model Pair 2"



(f) Model Pair 3"

Figure F1: Vectors of Posteriors (Third study)

Notes. Panels (a) and (b) provide results for the *Baseline* condition, (c) and (d) for *Non-Conflicting* condition, and (e) and (f) for *Three-Models* condition. The size of the circles is relative to the number of observations. We pooled observations that are within 2 p.p. of the point predictions. Observations in the gray areas are Bayes-consistent, while observations in the white areas are Bayes-inconsistent.

Type of Guess	Baseline		Non-Conflicting		Three-Models	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	3.36	[0.92, 5.81]	2.09	[0.43, 3.75]	12.28	[7.96, 16.60]
One-stage	10.40	[6.24, 14.55]	14.29	[9.36, 19.22]	16.14	[11.04, 21.24]
Best-fitting	42.51	[36.18, 48.84]	41.46	[34.65, 48.28]	28.77	[23.00, 34.55]
Worst-fitting	6.42	[3.44, 9.40]	4.88	[2.42, 7.34]	2.46	[0.65, 4.26]
Middle model					12.28	[7.96, 16.60]
Within	27.52	[21.83, 33.21]	24.39	[18.84, 29.94]	30.88	[24.66, 37.10]
Outside	9.79	[6.15, 13.42]	12.89	[8.46, 17.32]	9.47	[5.72, 13.23]
Total	100.00		100.00		100.00	

Table F8: Classification of Guesses by Treatment, Model Pair 2, Exact (Third Study)

Notes. The table reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, worst-fitting, and middle model) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise) for each treatment (*Baseline*, *Non-Conflicting*, *Three-Models*). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We use only data from the updating tasks that are based on Model Pair 2. Note that, in *Three-Models*, 12.28% of reported guesses are classified as both Bayesian and middle model. These duplicate observations are not included in the total.

Type of Guess	Baseline		Non-Conflicting		Three-Models	
	%	95%-CI	%	95%-CI	%	95%-CI
Bayesian	1.40	[0.03, 2.78]	2.17	[0.17, 4.16]	2.02	[0.43, 3.61]
One-stage	8.07	[4.64, 11.50]	13.36	[8.70, 18.02]	14.48	[9.67, 19.28]
Best-fitting	42.46	[35.91, 49.00]	35.38	[28.94, 41.82]	26.60	[20.74, 32.46]
Worst-fitting	6.67	[3.34, 9.99]	10.83	[6.66, 15.00]	5.39	[2.36, 8.41]
Middle model					8.42	[4.94, 11.90]
Within	30.53	[24.56, 36.49]	27.44	[21.46, 33.41]	32.32	[26.18, 38.47]
Outside	10.88	[7.07, 14.68]	10.83	[6.54, 15.12]	10.77	[7.05, 14.50]
Total	100.00		100.00		100.00	

Table F9: Classification of Guesses by Treatment, Model Pair 3, Exact (Third Study)

Notes. The table reports the shares of guesses that can be exactly classified as one of the point predictions (Bayesian, one-stage, best-fitting, worst-fitting, and middle model) or in one of the two residual categories ("Within" if the guess is within the two model predictions or "Outside" otherwise) for each treatment (*Baseline*, *Non-conflicting*, *Three-Models*). Columns "%" and "95%-CI" report the shares and corresponding 95% confidence intervals, using standard errors clustered at the individual level. We use only data from the updating tasks that are based on Model Pair 3.

Type of Guess	Baseline vs. Non-Conflicting				Baseline vs. Three-Models			
	Coeff.	t	p	95%-CI	Coeff.	t	p	95%-CI
Bayesian	-0.32	-0.27	0.79	[-2.69, 2.05]	4.59	3.04	0.00	[1.62, 7.56]
One-stage	4.52	1.63	0.10	[-0.93, 9.97]	5.98	2.11	0.04	[0.41, 11.54]
Best-fitting	-4.01	-0.97	0.33	[-12.14, 4.12]	-14.82	-3.81	0.00	[-22.46, -7.18]
Worst-fitting	1.27	0.70	0.48	[-2.27, 4.81]	-2.58	-1.58	0.12	[-5.81, 0.64]
Model selection	-2.74	-0.61	0.54	[-11.54, 6.05]	-7.10	-1.61	0.11	[-15.74, 1.55]
Best-fitting, cond.	-3.52	-1.01	0.32	[-10.42, 3.37]	-20.68	-4.68	0.00	[-29.40, -11.97]
Identified	1.45	0.36	0.72	[-6.48, 9.38]	-2.54	-0.63	0.53	[-10.46, 5.39]

Table F10: Treatment Comparisons, Exact (Third Study)

Notes. This table presents treatment effects by regressing indicator variables, capturing whether participants are classified under different updating rules (requiring an exact match; see Table 9), on treatment dummies. The analysis pools data from all updating tasks and clusters standard errors at the individual level. Note that, in *Three-Models*, 6.01% of reported guesses are classified as both Bayesian and middle model. The column "Baseline vs. Non-Conflicting" compares the *Baseline* and *Non-Conflicting* conditions (N=1,176), while the column "Baseline vs. Three-Models" compares the *Baseline* and *Three-Models* conditions (N=1,176). For the row "Model Selection," the outcome variable is defined as 1 if the participant selects either the best- or worst-fitting model (or, in the *Three-Models* condition, the middle model) and 0 otherwise. For the row "Best-fitting, cond.," we restrict the sample to observations where participants engage in model selection (N=561 for *Non-Conflicting* and N=544 for *Three-Models*). For the row "Identified," the outcome variable is defined as 1 if the observation is classified as Bayesian, one-stage, best-fitting, worst-fitting, or middle model, and 0 otherwise.

Type of Guess	Baseline vs. Non-Conflicting				Baseline vs. Three-Models			
	Coeff.	t	p	95%-CI	Coeff.	t	p	95%-CI
Bayesian	-1.27	-0.85	0.40	[-4.22, 1.67]	8.92	3.54	0.00	[3.97, 13.87]
One-stage	3.89	1.19	0.23	[-2.54, 10.31]	5.74	1.72	0.09	[-0.81, 12.30]
Best-fitting	-1.04	-0.22	0.82	[-10.31, 8.22]	-13.74	-3.16	0.00	[-22.27, -5.20]
Worst-fitting	-1.54	-0.79	0.43	[-5.40, 2.31]	-3.97	-2.25	0.03	[-7.44, -0.49]
Model selection	-2.59	-0.53	0.60	[-12.22, 7.05]	-5.42	-1.12	0.26	[-14.92, 4.08]
Best-fitting, cond.	2.60	0.66	0.51	[-5.12, 10.32]	-20.75	-3.91	0.00	[-31.22, -10.27]
Identified	0.03	0.01	1.00	[-8.98, 9.03]	-3.04	-0.67	0.50	[-12.00, 5.92]

Table F11: Treatment Comparisons, Model Pair 2, Exact (Third Study)

Notes. This table presents treatment effects by regressing indicator variables, capturing whether participants are classified under different updating rules (requiring an exact match; see Table F8), on treatment dummies. The analysis pools data from the updating tasks based on Model Pair 2 and clusters standard errors at the individual level. Note that, in *Three-Models*, 6.01% of reported guesses are classified as both Bayesian and middle model. The column "Baseline vs. Non-Conflicting" compares the *Baseline* and *Non-Conflicting* conditions (N=1,176), while the column "Baseline vs. Three-Models" compares the *Baseline* and *Three-Models* conditions (N=1,176). For the row "Model Selection," the outcome variable is defined as 1 if the participant selects either the best- or worst-fitting model (or, in the *Three-Models* condition, the middle model) and 0 otherwise. For the row "Best-fitting, cond.," we restrict the sample to observations where participants engage in model selection (N=561 in *Non-Conflicting* and N=544 in *Three-Models*). For the row "Identified," the outcome variable is defined as 1 if the observation is classified as Bayesian, one-stage, best-fitting, worst-fitting, or middle model, and 0 otherwise.

Type of Guess	Baseline vs. Non-Conflicting				Baseline vs. Three-Models			
	Coeff.	t	p	95%-CI	Coeff.	t	p	95%-CI
Bayesian	0.76	0.62	0.54	[-1.66, 3.18]	0.62	0.58	0.56	[-1.48, 2.72]
One-stage	5.29	1.80	0.07	[-0.48, 11.05]	6.41	2.14	0.03	[0.53, 12.29]
Best-fitting	-7.08	-1.52	0.13	[-16.23, 2.08]	-15.86	-3.56	0.00	[-24.61, -7.10]
Worst-fitting	4.16	1.54	0.12	[-1.15, 9.48]	-1.28	-0.56	0.57	[-5.76, 3.20]
Model selection	-2.91	-0.59	0.56	[-12.61, 6.79]	-8.72	-1.77	0.08	[-18.38, 0.95]
Best-fitting, cond.	-9.87	-1.87	0.06	[-20.29, 0.56]	-20.60	-3.52	0.00	[-32.12, -9.07]
Identified	3.14	0.67	0.50	[-6.03, 12.30]	-1.69	-0.36	0.72	[-11.01, 7.62]

Table F12: Treatment Comparisons, Model Pair 3, Exact Match (Third Study)

Notes. This table presents treatment effects by regressing indicator variables, capturing whether participants are classified under different updating rules (requiring an exact match; see Table F9), on treatment dummies. The analysis pools data from the updating tasks based on Model Pair 3 and clusters standard errors at the individual level. Note that, in *Three-Models*, 6.01% of reported guesses are classified as both Bayesian and middle model. The column "Baseline vs. Non-Conflicting" compares the *Baseline* and *Non-Conflicting* conditions (N=1,176), while the column "Baseline vs. Three-Models" compares the *Baseline* and *Three-Models* conditions (N=1,176). For the row "Model Selection," the outcome variable is defined as 1 if the participant selects either the best- or worst-fitting model (or, in the *Three-Models* condition, the middle model) and 0 otherwise. For the row "Best-fitting, cond.," we restrict the sample to observations where participants engage in model selection (N=561 in *Non-Conflicting* and N=544 in *Three-Models*). For the row "Identified," the outcome variable is defined as 1 if the observation is classified as Bayesian, one-stage, best-fitting, worst-fitting, or middle model, and 0 otherwise.

G Appendix: Experimental Instructions & Interface

The instructions and screenshots of the experimental interface for all studies are available by clicking here: