Beliefs from Cues

John J. Conlon   Spencer Y. Kwon

March 27, 2025*

Abstract

We develop a model of belief formation in which normatively irrelevant cues influence which states come to mind. Beliefs depend on two key factors: representativeness, the similarity of a cue to high-utility states, and resonance, the cue's ability to evoke many states. We show how to derive testable predictions from the model from subjective similarity judgments, a method we showcase in a controlled experiment in which participants estimate probabilities in a task involving uninformative cues. Beliefs systematically shift with cues' (separately measured) resonance and representativeness, consistent with the model's predictions. Finally, we explore economic applications of our framework by applying our model to a standard consumption savings problem and risky investment. We show how cue-based beliefs can underlie (and make novel predictions about) seemingly disparate behavioral phenomena, such as experience effects, projection bias, present bias, uninformative advertising, and portfolio under-diversification.

[*] Conlon: Carnegie Mellon University, jconlon@andrew.cmu.edu. Kwon: Brown University, spencer_kwon@brown.edu. We thank Ned Augenblick, Francesca Bastianello, Pedro Bordalo, Andrew Caplin, Nicola Gennaioli, Matt Gentzkow, Thomas Graeber, Alex Imas, Botond Koszegi, Yueran Ma, Rafael La Porta, George Loewenstein, Yueran Ma, Muriel Niederle, Ryan Oprea, Chris Roth, Suproteem Sarkar, Josh Schwartzstein, Jesse Shapiro, Andrei Shleifer, Jason Somerville, Adi Sunderam, Ansgar Walther, Sevgyi Yuksel, and seminar audiences at the Bravo Center Conference for Behavioral Economics, SITE Experimental, the Bocconi Cognitive Foundations of Finance Conference, HBS Cognitive Economics Workshop, NYU Behavioral Seminar, Bonn Behavioral Seminar, and the IZA Beliefs Workshop for helpful comments and feedback. The experiment presented in this paper was pre-registered on the AEA registry (AEARCTR-0015484).

1 Introduction

Standard models in economics—even those incorporating behavioral frictions—tend to assume that beliefs are a function only of objectively relevant pieces of information, such as an agent's priors and the likelihood of signals. At the same time, there is a widespread understanding that this picture must be incomplete: individual expectations seem overly sensitive to immediate circumstances and personal experiences; advertisements include elaborate (but seemingly irrelevant) contextual details to help customers imagine enjoying a product; and researchers measuring expectations worry that survey framing and question wording may alter beliefs by distorting what considerations come to mind.[1] In other words, beliefs appear sensitive to normatively irrelevant *cues*.

While this fact is widely understood (at least implicitly), economists have largely avoided incorporating cues into formal models,[2] perhaps out of concern that it cannot be done in a way that retains the precision and falsifiability of standard economic theory. After all, how exactly do cues work? What precisely do they bring to mind, and what form (and what magnitude) of belief biases therefore arise? And can a theory of cues discipline the answers to these questions in a data-driven way? In this paper, we address these questions head-on.[3] Drawing on past work (Bordalo et al., 2023), we first formalize a simple model of beliefs from cues and show that it yields falsifiable predictions as well as clear guidance about the data required to test them. In a controlled experiment, we then show that the model succeeds in explaining belief distortions in response to hundreds of different cues, despite i) having very few degrees of freedom, and ii) standard rational and behavioral economic models predicting null effects. Finally, we show that embedding our model into otherwise canonical frameworks helps organize existing behavioral anomalies and generates novel economic predictions.

In our model, an agent decides whether to take an action whose value depends on

[1] See Taubinsky et al. (2024) for evidence on expectations being excessively sensitive to local contexts. The fact that many advertisements appear to lack informational content has long been documented (Stern & Resnik 1991) and studied by economists (e.g, Milgrom & Roberts 1986, Becker & Murphy 1993). For reviews of experimental/survey design, see Haaland et al. (2023) and Stantcheva (2022), each of which discusses the hazards that framing poses for researchers.

[2] There are, naturally, some important exceptions. Our model builds off of a recent literature studying the role of similarity, retrieval, and simulation in economics (e.g., Mullainathan, 2002; Mullainathan et al., 2008; Bordalo et al., 2023; Bordalo, Gennaioli, et al., 2024; Graeber et al., 2023; Enke, Schwerter, & Zimmermann, 2024; Bohren et al., 2024). See the literature review for more discussion.

[3] For example, Rabin (2013) writes, "The aim to have realism-improving theories be maximally useful to core economic research suggests a particular approach... One should (i) extend the existing model by formulating a modification that embeds it as parameter values with the new psychological assumptions as alternative parameter values, and (ii) make it portable by defining it across domains using the same independent variables in existing research, or proposing measurable new variables." This is precisely what the present paper aims to do for cues.

the state of the world. For example, the value of saving today depends on the agent's belief about the expenditures she will face in the future. A standard rational agent would simply integrate over all possible future states to estimate the action's expected utility. Our agent, instead, must rely on which states come to mind and, crucially, is more likely to think of states that are *similar* to the context she is forming her belief in. For example, a person might feel a greater need to save for retirement while visiting a grandparent at an expensive assisted-living facility. Or a consumer may fail to save for unusual or rare expenses like health emergencies—which are dissimilar from her typical context—absent an intervention explicitly prompting her to think about them (Augenblick et al. 2023). We broadly call these external features of the agent's environment "cues," and study how they can alter beliefs by shifting which states come to mind.

When the context consists of a single cue, we show that belief distortions depend on what we call (borrowing from Kahneman & Tversky 1972) its *representativeness*: how similar the cue is to states where the action yields high utility.[4] When the context contains multiple cues, beliefs distortions are a convex combination of the representativeness of each component cue. The weight a particular cue receives depends on its *resonance*: that is, how many states are similar to it and therefore come to mind when it is part of the agent's context.

We then show how to derive simple, testable comparative statics from the model. Predicting the effect that a cue will have on beliefs requires estimates of its resonance and representativeness, which ultimately depend on the underlying similarity between states. One might worry that, if similarity is unobservable, then any pattern of beliefs can be rationalized with *post hoc* assumptions. However, although economists typically do not collect data on similarity, cognitive psychologists have long elicited and analyzed data on subjective similarity judgments to map out associations (e.g. Tversky & Gati 1982, Ashby & Perrin 1988, Nosofsky 1992). With data on such similarity judgments, one can separately estimate the resonance and representativeness of any given cue and then test whether it affects beliefs as the model predicts.

We showcase this method, and thereby test the model, in a controlled experiment. Participants are tasked with forming beliefs about a simple data-generating process, framed as a "game." Each game consists of seven rounds that either add or subtract points from its total score. The number of contributed points each round is uniformly and independently drawn from the set $\{+1, +2, +3, +4, -5, -6\}$, framed as rolls of a six-sided die.

[4] When the utility from the action is continuous (rather than binary) across states, the representativeness of the cue is formally defined as the normalized covariance between how similar the cue is to various states and the utility of the action in those states.

Participants then form incentivized beliefs regarding the distribution of the final score, which is the sum across all seven rounds. Our primary question asks participants the probability that the game ends up with a final score greater than zero, though we also ask the likelihood of games with even vs odd final scores. Note that the large number of states—all sequences of seven die rolls—precludes participants from simply enumerating all possible outcomes, raising the possibility that cues may influence which states come to mind and thereby distort beliefs.

To test this possibility, participants are randomized to either form beliefs absent any additional cue or to do so only after watching an example game. These example games—the cues in our experiment—are objectively uninformative: the data-generating process is already fully described, logically pinning down the objective answers, and the example is only one of the hundreds of thousands of equally likely possible outcomes. Thus, the rational benchmark—and indeed the benchmark for any agent who responds only to objectively informative signals (e.g, even one who under- or overreacts to data)—is that these cues should have no effect on beliefs. Furthermore, participants are explicitly told—and must indicate their understanding—that the examples are randomly assigned and should not have any effect on their beliefs. After belief elicitation, participants are then asked to make pairwise similarity judgments between example games. We use these similarity data to estimate the resonance and representativeness of cues.

We find that these cues can have large effects on beliefs, and that these effects are systematically related to their resonance and representativeness (separately measured from similarity data). Moving from a cue that is representative of negative scores to one representative of positives scores increases beliefs about the share of positive outcomes from 39.8% to 58.9% (77% of a standard deviation, $p < 0.001$). Furthermore, cues that depict explicitly impossible states—ones that participants understand fail to follow the rules of the game—also have large effects on beliefs (54.6% vs 43.8% for positive vs negative cues, $p < 0.001$) because they are nonetheless similar to possible states. In contrast, beliefs about the share of games that end with even vs odd scores are not affected by cues. Our model successfully predicts this null result: participants do not judge even vs odd cues as differentially similar to even outcomes, and therefore these cues do not differ in how representative they are of that hypothesis. Finally, we elicit suggestive evidence on which outcomes participants felt came to mind while forming beliefs: these process data strongly track treatment effects on beliefs.

We then structurally estimate the model using average beliefs across 360 randomly generated cues. Despite having only three degrees of freedom, we find that our fitted model is able to closely track average beliefs across cues and across questions: it predicts the size and shape of effects of positive vs negative cues on beliefs about positive outcomes while also correctly predicting no effect of even vs odd cues. Following Fudenberg et al. (2022, 2023), we calculate the completeness and restrictiveness of the model: how well it fits our data relative to a best-case benchmark (completeness) and how poorly it spuriously fits synthetically generated fictitious data (restrictiveness). We find the model achieves 77.3% of the explanatory power of the best-case machine-learning benchmark while being 97.6% restrictive: that is, the model very well fits the actual data while being almost entirely unable to fit synthetic data. These results highlight how the model generates sharp, falsifiable predictions, which largely succeed in describing the empirical effect of cues.

Having provided experimental evidence for the model, we then show that beliefs from cues may help to unify and shed additional light on seemingly disparate behavioral anomalies. We start by embedding our theory into an otherwise standard model of intertemporal choice under uncertainty, where an agent decides how much to consume today versus save for the future. The agent's ability to imagine future states depends on whether they are cued by her context, which we assume includes the states she has already lived through (including the current state). We derive a similarity-augmented Euler equation that specifies how two factors distort the agent's savings decisions away from the rational benchmark: what we call differential cueing and cognitive discounting.

First, the agent's savings decision depends on whether her context differentially cues high marginal-utility future states. This leads the agent to exhibit both *experience effects* (Malmendier & Wachter 2022)—acting as though the future will resemble states she has personally lived through—and *projection bias* (Loewenstein et al. 2003)—acting as though the future will resemble today. If she has lived through high marginal-utility states (e.g., unemployment or the Great Depression), or if she is currently in such a state, she will save more and consume less (than the rational benchmark), even if these cues are not objectively informative about the distribution of future states. Critically, these forces operate through, and therefore are moderated by, similarity. Our framework therefore provides a method of determining the domain-specificity of these effects, which can disappear or even reverse sign when the agent is faced with dissimilar risks. For example, an older individual who lived through the Depression might be more cautious in financial domains (Malmendier & Nagel, 2011) but less so regarding health risks (Bordalo, Burro, et al., 2024).

The preceding discussion highlights how when an agent forms beliefs about the future, states similar to her context disproportionately come to mind. However, the extent to which future time periods at different horizons come to mind *at all* depends on strongly her context cues them: that is, on how similar on average the states within those periods tend to be to the present. This observation provides a novel link between projection/experience effects on the one hand (reflecting *which* future states come to mind) and apparent time discounting on the other hand (reflecting how *many* future states come to mind). In particular, our agent will look more patient when her present circumstances are similar to many future states but more imprudent in unusual circumstances: she will scrimp at the grocery store but splurge while on a vacation. This same force could operate over the life cycle: an agent will become more patient over time the longer she has lived in situations that resemble the (likely) future. Thus a worker who has previously bounced between jobs or locations will slowly begin to save more once she lands a steady career in a city she likes (the opposite of what a standard model would predict given the reduction in uncertainty).

Further, on the intuitive assumption that the distribution of states is stationary, the agent endogenously exhibits dynamically-inconsistent *present bias* (Laibson 1997, O'Donoghue & Rabin 1999): her discount factor is especially severe over the immediate future but levels off over longer horizons. Similarity drives this result. The "dropoff" in similarity to the present (and therefore in ease of imagination) between today and tomorrow can be severe, as any unusual present circumstances likely will not persist. But the decline in similarity to the present between 365 days from now and 366 days from now will be negligible (and, in the limit, zero). Thus our agent is especially impatient regarding benefits that can accrue in the near future, but will trade off benefits between more distant dates (asymptotically) rationally.

So far, we have considered a case where an agent faces exogenous cues. Our experiment showed however that externally provided cues have large effects on beliefs, even in cases where agents are explicitly told they are normatively irrelevant. This raises the question of how a persuader, armed with the ability to cue strategically chosen states, can manipulate the agent to her advantage. To study this, we model a persuader that designs a product along with a cue to boost the agent's willingness to pay for it. For example, a financial advisor might both construct a portfolio of assets and then try to "sell" her client on it by cuing her with its upside. Or a firm might produce a consumer good—whose overall usefulness may depend on how often the customer finds herself in certain situations (e.g., sunny weather for a convertible car)—as well as an advertisement to help its customers imagine how they will use the product. We then explore the implications of this cueing technology on the seller's product design.

Our main result highlights a novel tradeoff between riskiness and *persuade-ability*. Consider a financial advisor designing a portfolio for a risk-averse client. A perfectly safe asset yields the same return in all states, but for precisely this reason it is impervious to cueing: there is no cue the advisor could provide that would boost the agent's beliefs about its expected return. In contrast, a risky portfolio (one, say, heavily invested in a particular industry) raises the possibility of persuasion through cueing ("The tech sector might take off!"). The client therefore ends up with an under-diversified portfolio, because it maximizes the fees the advisor can extract. In a consumer-product setting, this same force leads to greater *ex post* regret than in the rational benchmark. The firm designs more "specialized" products—like warranties or time-share vacation homes—with particular, high-utility use cases, which the firm cues in its advertisements to customers. Buyers in turn end up in these high-utility situations less than it felt like they would when they purchased the product. These analyses show how cue-based persuasion can affect more tangible aspects of the economy like portfolio compositions and product design.

Literature This paper contributes first to a literature on decision-making when agents cannot retrieve or process all relevant information (e.g., Bordalo, Gennaioli, et al. 2024, Woodford 2020, Caplin et al. 2019, Gabaix 2019, Enke & Graeber 2023, Bordalo et al. 2025, Graeber et al. 2023, Ba et al. 2023, Bohren et al. 2024, Oprea 2024, Conlon 2024, Augenblick, Backus, et al. 2024). Our paper also speaks to the related literature on failures of contingent reasoning (see Niederle & Vespa 2023 for a review), where agents fail to fully or accurately consider states in which their actions are relevant (e.g., Esponda & Vespa 2014, Martínez-Marquina et al. 2019, Moser 2019).[5] Some of this literature studies cues, though typically focusing on the role of cues in bringing *memories* to mind (Bordalo et al. 2023, Enke, Schwerter, & Zimmermann 2024, Bordalo, Gennaioli, et al. 2024). Our work shows that a similar sampling framework can be productively applied to also study which *unrealized* future states come to mind when agents form beliefs (see Bordalo, Burro, et al. 2024 for related, and complementary, ideas).

Our work also complements a growing literature on changing mental models (Eliaz & Spiegler 2020, Schwartzstein & Sunderam 2021, Barron & Fries 2023, Charles & Kendall 2024). Existing theories tend to assume that, of the models people are exposed to, they adopt only one (e.g., the best-fitting model or the one that delivers the highest anticipatory utility). Our focus instead is on associations: our agent equally "counts" all states

[5] An interesting difference between our framework and the literature on failures of contingent reasoning is that these papers often focus on settings with few (e.g., only two) relevant states; we speculate that one force driving some failures of contingent reasoning in these studies may be the difficulty of correctly or fully simulating a given state. For example, participants in Esponda & Vespa (2014)'s experiment appear not to realize absent prompting that there is *no* state where one action out-earns the alternative action. We abstract away from this consideration in our study and instead focus on the question of *which* states come to mind, rather than the ability to accurately simulate a given state conditional on it doing so.

that she thinks of, but which states come to mind depends on the similarity between them and the cues she faces.

Next we contribute to a literature on simulation in economics and psychology. The role of simulation (the process of imagining future states/scenarios) in forming beliefs has long been recognized and studied in psychology (e.g., Kahneman & Tversky 1981, Dougherty et al. 1997, Gershman et al. 2017, Chater et al. 2020). It has received less attention in economics, with some important exceptions. For example, Bordalo, Burro, et al. (2024) study how memories of past experiences can aid or inhibit simulation of future risks. Becker & Mulligan (1997) model an agent who can invest in her ability to appreciate the future, while Gabaix & Laibson (2022) study how an agent who (realizes she) can only noisily approximate future utility flows will often seem to act impatient or present biased.

Lastly, as alluded to above, our model closely resembles theories of memory retrieval (Kahana 2012, Bordalo et al. 2023), reflecting work in psychology finding overlap between the brain processes underlying simulation and memory (Schacter et al. 2007, Gilbert & Wilson 2007, Schacter et al. 2012, Mullally & Maguire 2014, Benoit & Schacter 2015). Our focus on the importance of similarity between states in facilitating simulation echoes other work in economics studying similarity as a underlying driver of many cognitive processes, such as memory retrieval (Mullainathan 2002, Charles 2022b,a, Bordalo et al. 2023, Graeber et al. 2023, Link et al. 2023, Enke, Schwerter, & Zimmermann 2024, Bohren et al. 2024, Colonnelli et al. 2024), categorization (Mullainathan et al. 2008, Bordalo, Gennaioli, et al. 2024, Evers et al. 2021), reinforcement learning (Barberis & Jin 2023), and reasoning by analogy (Gilboa & Schmeidler 1995).

2 Model

In this section, we first describe the basic setup of the model, which formalizes how an agent constructs beliefs by sampling states of the world. States come to mind more readily when they are more similar to the cues the agent faces, leading to belief distortions. We show that two properties of cues—summarizing whether they bring many states to mind and the extent to which those states favor a particular hypothesis—determine both the extent of and the limits to their effects on beliefs. These lead to testable comparative statics based on the underlying similarity structure of the state space, which can be proxied for by eliciting similarity judgments.

2.1 Setup

Objective environment Let Ω be the set of possible states of the world, with the likelihood (or frequency) of $\omega \in \Omega$ given by $\pi(\omega)$. The agent is considering an action whose payoff $\psi(\omega)$ is state-dependent. Under the frictionless rational benchmark, the expected utility of the action is then given by $\Psi_{rat} \equiv \sum_{\omega\in\Omega} \pi(\omega)\psi(\omega)$.

Similarity The key idea that our model formalizes is that, unlike a standard rational agent, an agent may struggle to consider all possible states, and the context in which she forms her beliefs may influence which states come to mind. In particular, we assume that states are more likely to come to mind the more similar they are to the context the agent faces. This implies that normatively irrelevant elements of her context—which we call "cues"—can distort her beliefs by bringing similar things to top of mind. Formally, we assume an exogenous similarity function $S : \Omega \times \Omega \mapsto [0,1]$ on the state space, where $S(\omega,\omega) = 1$ for all ω. $S(\omega_1, \omega_2)$ captures the degree to which ω_1 is similar to (or "associated with") ω_2. We can then generalize S to apply to subsets of Ω by averaging element-wise similarity. Formally, let A and B be subsets of Ω. Then, the similarity between A and B is given by:

$$S(A,B) = \frac{1}{|A| \cdot |B|} \sum_{\omega\in A,\omega'\in B} S(\omega,\omega'). \tag{1}$$

Context and cued beliefs We primarily focus on cues that can be modeled as elements of Ω. Formally, we define the context C as a subset of Ω. We assume the agent forms beliefs about the expected utility Ψ_{rat} by trying to imagine (or "simulate") possible states. Drawing from past work (Bordalo et al., 2023), we model this as an associative sampling process. In particular, we assume the agent has an unbiased "mental database"—she is aware, in a sense, of all possible states ω and their objective likelihood $\pi(\omega)$—but the sampling process that allows her to access these states is distorted by the context C. In particular, we assume the probability she samples a state ω is proportional to its similarity to the context $S(\omega, C)$. In the large sample limit, the agent's simulated distribution is then proportional to:

$$\pi_s(\omega|C) \propto \pi(\omega) \cdot S(\omega, C), \tag{2}$$

where $S(\omega, C)$ is given by Equation 1. We assume that the agent is naive about the context distortion: her beliefs are thus given by the simulated distribution $\pi_s(\omega|C)$. Equation 2 highlights that the beliefs cued by the context can be simply viewed as the true distribution π with context-dependent distortion weights $S(\omega, C)$. When sampling is unbiased ($S(\omega, C) = 1$), one recovers the rational benchmark.

Interpretation We view our model as providing a tractable way of incorporating associative (i.e., similarity-based) cues into a model of belief formation. Naturally, many assumptions are simplifications. For example, the similarity function in our model is exogenous, but other work microfounds and endogenizes similarity by considering how many salient features two items have in common (Tversky, 1977; Bordalo et al., 2025). We also primarily model cues simply as subsets of the state space, which is restrictive: for example, our model will fail to incorporate effects of describing a state in different terms or of cues that differ greatly in type from states of the world (e.g., an agent's current emotions). But our framework provides a natural starting place for analyzing the effect of many common cues: for example, the agent's current situation, her experience of past states, depictions or descriptions of scenarios (e.g., in advertisements), and so on.

Finally, though for simplicity we describe our agent as making a choice under uncertainty—where she is evaluating the probability of different states of the world—our framework can be applied more generally. The crucial ingredients are that there be 1) a large set of objects about which agents are forming beliefs, and 2) an *associative structure* over these objects. Though risky choice under unknown states of the world is a natural example of a setting with these two inputs, there are others that lack any "true" uncertainty. For example, an agent may be forming beliefs about her own past experiences, where her recall of memories is partly shaped by their associations (Bordalo et al., 2023). Or an agent may be trying to evaluate how often she will find herself in various situations in the future, about which she may in principle already know enough to decide with certainty. All of the analysis to follow will equally apply even to these cases where there is no "true" uncertainty.

2.2 Characterizing cued beliefs

Characterizing beliefs To build intuition, we first focus on the case where $\psi(\omega) = I(\omega \in H)$, for $H \subset \Omega$. In that case, $\Psi_{rat} = \pi(H)$: the agent is assessing the likelihood of a hypothesis H, which is a subset of the state space. Equation 2 implies the following lemma regarding how a given context $C \subset \Omega$ distorts the assessment of $\pi(H)$.

Lemma 1. *Let $\mathcal{W}(\omega_c) = E[S(\omega,\omega_c)] = \sum_{\omega\in\Omega} \pi(\omega)\cdot S(\omega,\omega_c)$ be the average similarity of ω_c to $\omega \in \Omega$ and $\mathcal{D}^H(\omega_c) = \frac{\sum_{\omega\in H} S(\omega,\omega_c)\pi(\omega)}{\sum_{\omega\in\Omega} S(\omega,\omega_c)\pi(\omega)} - \pi(W)$. The agent's belief given context C is:*

$$\underbrace{\pi_s(H|C) - \pi(H)}_{\textit{Context distortion}} = \frac{\sum_{\omega_c\in C} \mathcal{W}(\omega_c)\cdot\mathcal{D}^H(\omega_c)}{\sum_{\omega_c\in C} \mathcal{W}(\omega_c)}. \tag{3}$$

Lemma 1 shows that the effect of $\omega_c \in C$ on the assessment of $\pi(W)$ is characterized by two summary statistics based on similarity. The first is $\mathcal{W}(\omega_c) = E[S(\omega, \omega_c)]$, which we call the *resonance* of ω_c, the average similarity of ω_c to other states in Ω. Because cues bring states to mind in proportion to their similarity, the resonance of ω_c captures the extent to which ω_c brings many states to mind. The second quantity, $\mathcal{D}^H(\omega_c)$, is what we call the *representativeness* of ω_c with respect to hypothesis H. $\mathcal{D}^H(\omega_c)$ intuitively captures whether the states that ω_c bring to mind tend to be consistent with hypothesis H. Equation 3 implies that the effect of context C on the agent's beliefs about H is a convex combination of the representativeness of each of its constituent cues, with weights proportional to their resonance.

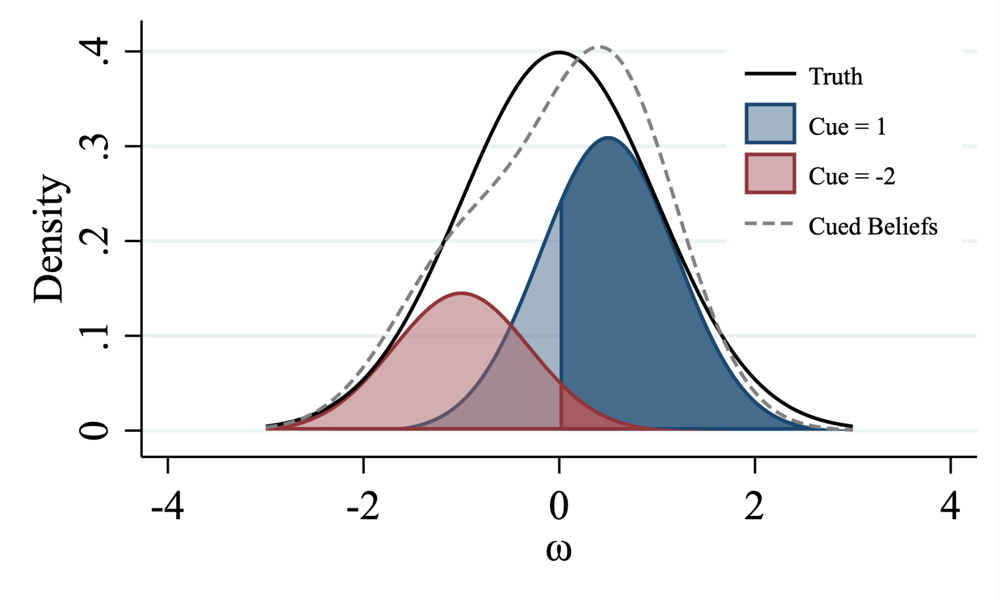


Figure 1: Context-distorted beliefs

Figure 1 visualizes the effect of a context containing two cues on the agent's beliefs. We consider the case in which $\Omega = \mathbb{R}$, with the true distribution π given by the normal distribution $\mathcal{N}(0,1)$. We endow Ω with the similarity function $S(\omega, \omega') = \exp\left(-\frac{\kappa}{2}(\omega - \omega')^2\right)$. Suppose that the agent is assessing the likelihood of $H = \{\omega \geq 0\}$, where the context consists of two cues: $C = \{-2, 1\}$. The red and blue curves in Figure 1 plot respectively $S(\omega, 1) \cdot \pi(\omega)$ and $S(\omega, -2) \cdot \pi(\omega)$, the beliefs that each cue would induce absent the other. The resonance of each cue is given by the total mass under each curve, while the representativeness of each cue with respect to H is given by the ratio between the darker and lighter areas for each curve (minus the true $\pi(H)$). As the figure illustrates, $\omega_c = 1$ both has higher resonance and representativeness (with respect to H) than $\omega_c = -2$. The resulting context-distorted beliefs are shown by the dashed curve.

Proposition 1 generalizes the expression in Lemma 1 to any utility function $\psi(\omega)$:

Proposition 1. *The agent's beliefs $E_s[\psi(\omega)|C]$ in context C are:*

$$\overbrace{E_s[\psi(\omega)|C] - E[\psi(\omega)]}^{\textit{Effect of cues}} = \frac{\sum_{\omega_c \in C} \mathcal{W}(\omega_c) \cdot \mathcal{D}^{\psi}(\omega_c)}{\sum_{\omega_c \in C} \mathcal{W}(\omega_c)}, \tag{4}$$

where $\mathcal{D}^{\psi}(\omega_c) = \frac{Cov[S(\omega,\omega_c),\psi(\omega)]}{E[S(\omega,\omega_c)]}$ is the representativeness of ω_c with respect to ψ.

Note that the only difference between the expressions in Lemma 1 and Proposition 1 is in the definition of representativeness. Representativeness in Proposition 1 is defined as the covariance between the similarity to ω_c and the value of ψ. Thus, it captures the representativeness of ω_c with respect to high-utility states. The function is equivalent to the previous notion of representativeness when $\psi(\omega) = I(\omega \in H)$.

Possible and impossible cues While we have focused on the case where our cues correspond to possible states of the world $\omega_c \in \Omega$, our framework regarding the effect of cues (and equation 4) generalizes to any object $\bar{\omega}_c \in \bar{\Omega}$ that has a well-defined similarity relationship with elements in Ω. Of particular interest is the case in which $\Omega \subset \bar{\Omega}$, where elements $\bar{\omega} \in \bar{\Omega} - \Omega$ are states of the world that the agent knows to be impossible. Even then, cuing an agent with a scenario she knows to be impossible may still have an impact on her beliefs, because it can still distort what comes to mind. For example, many advertisements depict scenarios that are far-fetched for typical consumers (e.g. an SUV cruising through the wilderness), or an agent may be recalling a previous experience that she knows cannot exactly re-occur, but each of these cues can nonetheless bring to mind possible future scenarios. More broadly, note that the resonance of a cue *is not its likelihood*: an extreme or even impossible cue ($\pi(\omega_c) = 0$) may still have substantial resonance $\mathcal{W}(\omega_c)$ and sway beliefs if it brings to mind many similar states. This possibility also emphasizes how cues shape beliefs differently than informative signals: impossible cues may shift beliefs in predictable ways by shaping what comes to mind, whereas the rational update from an impossible signal is not well-defined.

2.3 Limits to cues

While our framework predicts that beliefs are malleable to cues and therefore to irrelevant changes in context, it also predicts limits to such effects. Not any beliefs can be achieved by providing the appropriate cues. In particular, because belief distortions are a convex combination of the representativeness of cues, the maximal distortion is bounded above by the maximal representativeness, as shown in equation 5:

$$E_s[\psi|C] - \Psi_{rat} \leq \max_{\omega \in C} \mathcal{D}^{\psi}(\omega). \tag{5}$$

Which beliefs are susceptible to cueing? Note that the maximal representativeness depends both on the underlying similarity function $S(\omega,\omega')$ and on $\psi(\omega)$, the utility function the agent is considering. Intuitively, if $\psi(\omega)$ and $S(\omega,\omega')$ are unrelated—states with high values of ψ are no more similar to each other than to states with low values—the maximal representativeness and thus scope for cue-based distortions will be small.

As an example, consider the following similarity structure, drawn from Tversky (1977). An element of $\omega \in \Omega$ is defined by a finite set of K features $\omega = (f_1(\omega), f_2(\omega), ...f_K(\omega))$, with $S(\omega,\omega') = \prod_{k=1}^{K} \delta_k^{I(f_k(\omega) \neq f_k(\omega'))}$. The parameter δ_k captures whether feature f_k is salient or not: the more people attend to f_k, the lower the δ_k – states that disagree on a salient feature are judged as less similar. For simplicity, suppose only one binary feature $f_1 \in \{0,1\}$ is salient, with $\delta_1 < 1$ and $\delta_k \approx 1$ for $k \neq 1$. Denote π_i, $i \in \{0,1\}$ as the frequency of $f_1 = i$ in Ω. Suppose an agent is assessing the likelihood of $H \subset \Omega$, and assume without loss of generality that $f_1 = 1$ is relatively more frequent in H than in its alternative: $\pi_{1|H} \equiv \pi(f_1 = 1|\omega \in H) \geq \pi_1$. Then the upper bound on the maximal representativeness of a cue with respect to H, which is also then an upper bound to the strength of context distortion, is as follows:

$$\pi_s(H|C) - \pi(H) \leq \max_{\omega} \mathcal{D}^{H}(\omega) = (1-\delta_1)\frac{\pi_{1|H} - \pi_1}{\delta_1 + (1-\delta_1)\pi_1} \cdot \pi(H). \tag{6}$$

As equation 6 immediately implies, $\max_{\omega} \mathcal{D}^{H}(\omega)$ is strictly increasing in $\pi_{1|H}$ and decreasing in π_1 and δ_1. That is, the more strongly correlated H is with the similarity-relevant feature f_1, the more strongly one can distort beliefs toward H by cueing the agent with $\omega \in \{f_1(\omega) = 1\}$. If instead H uncorrelated with the similarity-relevant feature, the beliefs about H are entirely impervious to cueing: there is no cue that differentially brings to mind states in H.

Aggregation limits Finally, note that equation 5 holds even in the limit as $|C| \mapsto \infty$: that is, even as the number of cues the agent is faced with grows. Thus, even many cues that individually would increase beliefs relative to the default may in combination have only a relatively muted effect. The reason is that while cues move beliefs by shifting which states are more available and top of mind, multiple cues also *interfere* with each other:

making one state more top of mind crowds out other states.

2.4 Testable predictions

Though the theory described above is simple, it contains two objects—the agent's context C and the similarity function S—that we have intentionally left quite general. The context might contain many cues: the agent's current situation, past experiences she is recalling, descriptions/depictions of future scenarios, or any items in her environment that bring particular possibilities to mind. And many features of these cues might in turn influence their similarity to states of the world and therefore their effect on the agent's beliefs. One might worry that such generality leaves the model without testable predictions: any pattern of beliefs could in principle be rationalized by postulating an appropriate cue and similarity function. What, then, is the empirical content of the theory?

Fortunately, two observations allow us to formulate testable predictions. The first is that, while the similarity between cues and states might be multifaceted, it is measurable. Decades of research in cognitive psychology has studied associations by analyzing subjective similarity assessments elicited from experimental participants (e.g., Nosofsky 1988, Tversky 1977), and recent work in economics has begun to adopt these insights (Bordalo et al. 2025). Furthermore, one need not map out the entirety of the similarity function S; Proposition 1 makes clear that it is sufficient to estimate the resonance and representativeness of cues.

Second, while the context C might contain many cues—such that identifying them all and characterizing their resonance/representativeness would be infeasible—we can nonetheless derive comparative statics regarding the effect of a *marginal* cue. This situation is familiar to economists: for example, we cannot observe the entire bundle of amenities that a job offers its employees, but we can derive predictions regarding the effect that a change in one of them (e.g., the wage) ought to have on worker behavior. Formally, suppose we wish to predict the effect that adding a cue Q to the context C will have on the agent's beliefs, where we assume for simplicity that the marginal cue does not overlap with her existing context ($C \cap Q = \emptyset$).[6] Equation 2 then implies:

$$E_s[\psi(\omega)|C \cup Q] - E_s[\psi(\omega)|C] = \frac{\mathcal{W}(Q)}{\mathcal{W}_C + \mathcal{W}(Q)}(\mathcal{D}^{\psi}(Q) - D^{\psi}_{C}) \tag{7}$$

[6] Our assumption holds almost surely if $|\Omega|$ is continuous. If one extends our theory such that C is a weighted subset of Ω, our theory can readily account for overlapping and repeated cues. Formally, we can define the context C as a weighted subset of Ω, with similarity between two weighted subsets of Ω, A and B, defined as: $S(A,B) = \frac{1}{|A|\cdot|B|}\sum_{\omega,\omega' \in \Omega} w_A(\omega) w_B(\omega') \cdot S(\omega,\omega')$, where $w_A(\omega), w_B(\omega)$ are the weight of ω in A and B. and $|A| = \sum_{\omega} w_A(\omega)$ and $|B| = \sum_{\omega} w_B(\omega)$. This allows one to allow for overlapping C and Q, model the effect of repeated cues, and allow for the presence of differentially salient elements of the context.

The above equation emphasizes that the effect of adding Q depends only on the relative resonance and representativeness of Q compared to the residual context C. If Q has a higher representativeness than C, adding it to the context will boost beliefs, but the extent of this change will depend on how resonant Q is compared to C.

Testing such predictions requires estimates of the resonance and representativeness each cue Q among a set of possible cues $\mathbb{Q}$. These can be computed using the Monte Carlo method. Suppose we have data on pairwise similarity judgments between Q and N random samples $\omega_{1\leq k\leq N} \sim \pi(\omega)$. We can then compute for each $Q \in \mathbb{Q}$:

$$\widehat{\mathcal{W}(Q)} = \frac{1}{N}\sum_{n=1}^{N} S(\omega_k, Q) \qquad \widehat{\mathcal{D}^{\psi}(Q)} = \frac{\sum_{n=1}^{N} S(\omega_n, Q)\psi(\omega_n)}{\sum_{n=1}^{N} S(\omega_n, Q)} - \Psi_{rat}.$$

With such estimates of $\mathcal{W}(Q)$ and $\mathcal{D}^{\psi}(Q)$ in hand, we can then test simple comparative statics. For example, for cues with similary $\mathcal{W}$,thosecues with a higher representativeness $\mathcal{D}^{\psi}$ should increase beliefs by more.

To summarize, one can derive falsifiable predictions from the model about the effect that each of a set of cues $\mathcal{Q}$ will have on beliefs by eliciting appropriate similarity data. In the next section, we employ this procedure in a simple lab-experimental environment as a proof of concept. We show that objectively uninformative cues can have large effects on beliefs that are predictable given separately elicited similarity judgments.

3 Experiment

We test the model's predictions in a simple controlled environment. Experimental participants form beliefs about what the final score of a "game" tends to be. We then test the impact of providing cues—highlighting particular outcomes of the game by having participants watch them occur—on participants' beliefs. We use similarity data elicited from participants to generate predictions for the effects of these cues, which we then test.

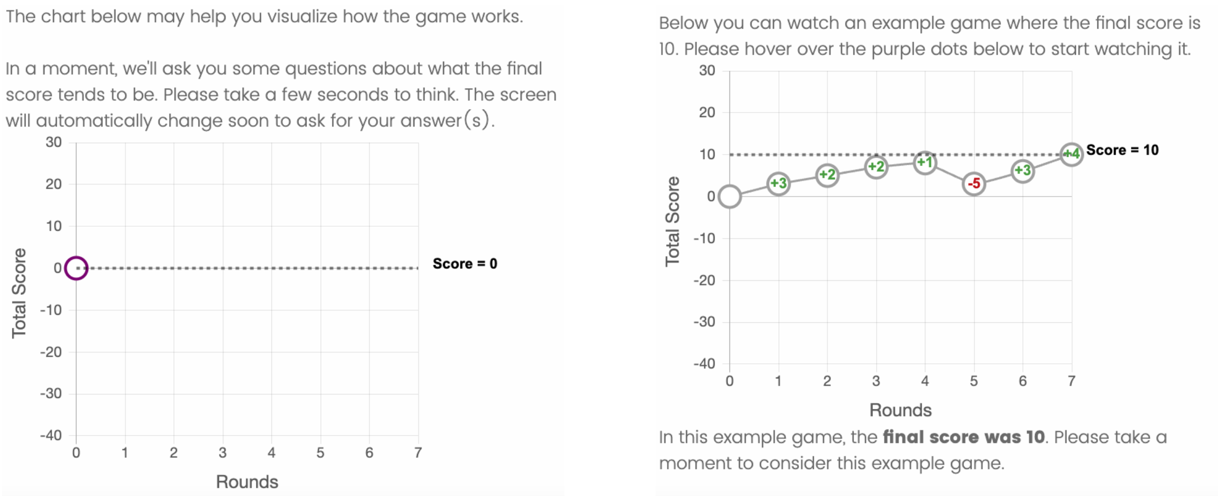
3.1 Design

The Environment. The experiment begins by describing to participants a simple data-generating process framed as a "game." Each game has seven rounds r, in each of which a number of points p_r is added to the "total score" of the game, where $p_r \in \{+1, +2, +3, +4, -5, -6\}$ with each outcome being equally likely and i.i.d. across rounds. To make this process intuitive for participants, it is described as rolling a six-sided die. The faces showing one through four are displayed in green, indicating that this number is added to the score, while five and six are shown in red, indicating subtraction. Participants are first tasked with guessing the share of games that end with a total score (adding up across the seven rounds) greater than zero. Afterward, they are also asked what fraction of games end up with an even total score (e.g., -4, -2, 0, 2, 4, ...).

Note that this environment maps directly onto the model described above. In the experiment, the set of states Ω is $\{+1, +2, +3, +4, -5, -6\}^7$ where each state corresponds to one of the 6^7 possible sequence of seven dice rolls. Further, these states are naturally grouped into focal hypotheses $H_p = \{s : \sum_{r=1}^{7} s_r > 0\}$ and $H_e = \{s : \sum_{r=1}^{7} s_r \text{ is even}\}$, the set of positive and even final scores, respectively. There are also a large number of states—6^7, or 279,936, of which 46.2% have a positive final score and 50.0% have an even final score—obviously precluding the strategy of simply enumerating all states consistent or inconsistent with the hypotheses. If participants form their beliefs by relying on which outcomes come to mind easily, their beliefs may be susceptible to cues.

Cues. After learning the rules of the game, all participants are shown a graph to help them visualize how it works, as shown in the left panel of Figure 2. Those in the control group, who are shown no additional cue, must simply wait 20 seconds before a text box appears asking them the average number of games per 100 that end up with positive scores and then with even scores. The remainder of participants are shown either one, five, or ten additional cues before the experiment elicits their beliefs. These cues (as depicted in the right panel of Figure 2) are framed as example games that participants watch unfold by hovering their mouse over dots that sequentially display the results of each round of the game. Participants must wait 20, ten, or five seconds after seeing each cue (for those seeing one, five, or ten cues, respectively) before either seeing an additional cue or before the text boxes appear eliciting their beliefs.

Note that, for several reasons, these cues are objectively (and explicitly) uninformative. First, the problem participants are solving does not involve any "true" uncertainty: the correct answers to the questions they are answering are (though difficult to ascertain) logically pinned down by the description of the data-generating process, which is fully specified even absent any example. Second, all outcomes are equally likely, so participants cannot infer that some outcomes are more likely from the fact that they were chosen to be shown as examples. And finally, we explicitly (and truthfully) tell all participants who receive a cue that we are choosing randomly whether to show them a cue and, if so, the type of cue to show them. Thus there is no objective signal to be extracted from the example game that a participant happens to see. Further, we explicitly state that "any example(s) you are shown should not impact your guesses", and participants must confirm their understanding of this fact in a comprehension question before continuing with the experiment. 95.4% of participants get this comprehension question correct on the first try, indicating a very high level of understanding that cues should be irrelevant.


The chart below may help you visualize how the game works.

In a moment, we'll ask you some questions about what the final score tends to be. Please take a few seconds to think. The screen will automatically change soon to ask for your answer(s).

Below you can watch an example game where the final score is 10. Please hover over the purple dots below to start watching it.

In this example game, the **final score was 10**. Please take a moment to consider this example game.

Figure 2: Belief elicitation with and without a cue

Similarity ratings. What effect will these cues have on participants' beliefs? The answer according to our framework depends on how similar they are different outcomes, and therefore which states come to mind in response to them. We borrow from a literature in psychology studying similarity (e.g., Tversky & Gati 1982, Ashby & Perrin 1988, Nosofsky 1992) and elicit similarity judgments directly from participants. More precisely, after providing their beliefs, all participants (regardless of treatment group) made 15 pairwise similarity judgments between outcomes of the game. Figure A.I shows what these elicitations looked like. Each pair of cues that participants compared included one cue from the set of cues shown to participants (described more below) an example game randomly generated from the true data-generating process. This procedure allows us to generate estimates of the resonance of each cue and its representativeness with respect to any hypothesis, using the Monte Carlo method described in Section 2.4.

Logistics and comprehension. A total of 3,154 participants, recruited from Prolific, completed our experiment in March 2025. The median participant took 12.5 minutes to complete the experiment and earned a $2.00 completion payment. Participants earned an additional $1.00 bonus if their belief about the fraction of positive or even games (decided randomly) was within 5 percentage points of the truth. The experimental instructions included ten comprehension questions, which participants had to answer correctly before they could continue with the experiment. Participants were truthfully told that those who gave more than two incorrect answers to comprehension questions would immediately be screened out of the experiment. This restriction only screened out 3% of subjects who attempted to participate in the experiment, indicating a high level of engagement and comprehension. In the main text, we restrict analysis to the 77% of the remaining participants who gave no incorrect answers to any of the ten comprehension questions, though we show in Appendix A that none of our main results are sensitive to this choice. Screenshots of the experiment, which was conducted through a Qualtrics survey, can be found in this online document. An interactive version of the experiment itself can be viewed/taken (with the ability to choose treatment assignment) by following this link.

3.2 Reduced-form results

Baseline positive and negative cues We begin by focusing on the reduced-form effect of a set of (pre-registered) "baseline" cues. These cues were selected as follows: we randomly selected ten outcomes each from the true data-generating process from among outcomes with final scores equal to -13, -12, -11, -10, +10, +11, +12, and +13. This process resulted in 80 baseline cues, the final scores of which are half positive, half negative, half even, and half odd. We can then compare the beliefs of the 309 participants assigned to see one of these cues to the 161 participants assigned to the control group (no cue).

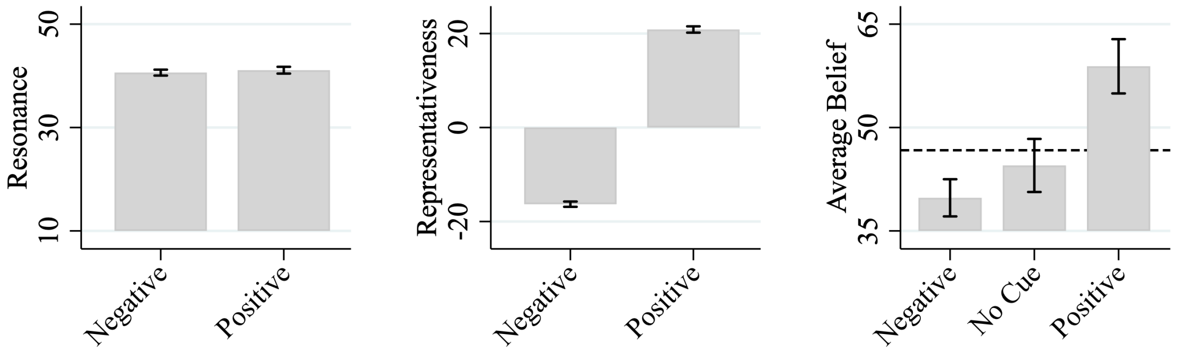
Figure 3 splits these baseline cues by whether their final score is positive or negative. For each group, it shows average resonance, representativeness, and beliefs about the probability that the final score is positive. Three facts stand out. First, positive and negative baseline cues are approximately equally resonant ($p = 0.640$): that is, on average they are equally similar to the true distribution of possible outcomes. Second, however, positive baseline cues are 37.7 p.p. ($p < 0.001$) more representative of outcomes with a positive final score than negative baseline cues are. That is, the positive baseline cues are judged by participants as being more similar to the average outcome with a positive final score than the negative baseline cues are. This result is exactly what we would expect if participants tend to judge games as similar to each other when their final scores are close to one another (which we show in Table A.I, see Appendix A for more details).

Third, the final panel of Figure 3 shows that beliefs differ dramatically across these treatment groups. Participants who see a baseline positive cue believe that 58.9% of games end with a positive score, compared to 39.8% among participants who see a negative cue ($p < 0.001$). Average beliefs in both groups are (at least marginally) significantly different than beliefs in the control group (44.5%, different from baseline negative at

$p = 0.051$ and from baseline positive at $p < 0.001$).[7]

These effects are striking for two reasons. First, as discussed in greater detail above, these cues are (explicitly) objectively uninformative, and participants had confirmed their understanding of this fact on the page immediately prior to the belief elicitation. Second, these effects are quite large. The main effect of 19.1 p.p. represents 77% of the standard deviation of beliefs in the control group. To provide a benchmark against which to compare this effect size, we can look at updating in response to objectively informative signals from the literature. For example, in Möbius et al. (2022), participants receive a 75%-accurate binary signal about whether their IQ-test score was above the median. Beliefs about the chances of actually having an above-median performance are 48% of a standard deviation higher for those who receive a positive vs a negative signal.[8] Thus, our uninformative cues have substantial effects on beliefs even compared to quite precise objectively informative signals in other contexts.

Figure 3: Effect of Baseline Cues on Beliefs about P(Final Score > 0)



Notes: This figure shows average resonance and representativeness of our baseline positive and negative cues (left and middle panels, respectively), each computed from our similarity elicitations. The right panel shows average beliefs across participants shown these cues, as well as the control group, about the probability that the game ends with a positive score.

Impossible cues Recall from Section 2.2 that a cue need not depict an outcome that is one of the possible states of the world in order to influence beliefs. Even a cue that the agent knows cannot happen in the future (e.g., fictional or historical states) can shape beliefs by bringing to mind similar states that are in fact possible. Our experimental paradigm allows us to test this prediction by constructing what we call "impossible" cues.

[7] Note that, because the representativeness of the default context is unobserved, the model does not make unambiguous predictions about what average beliefs in the control group should be.

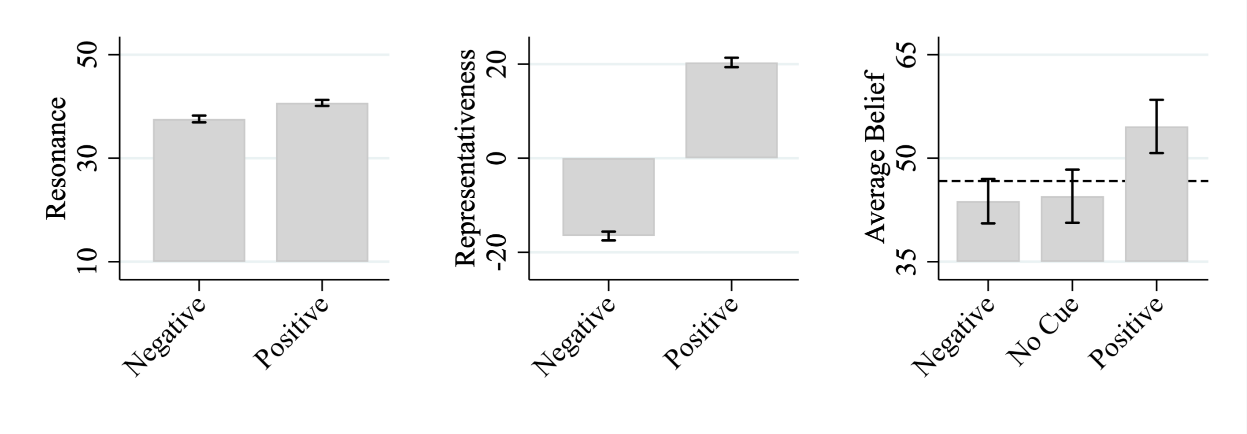
[8] This statistic comes from our own calculation using Möbius et al. (2022)'s replication files.

Recall that each round either adds 1, 2, 3, or 4 points, or it subtracts 5 or 6 points. We construct impossible states by showing outcomes that break these rules but otherwise follow the general structure of the data-generating process. In particular, we start by randomly picking baseline cues with a final score equal to either -13 or +10. We then randomly choose one round within these cues. If that round added points (+1, +2, +3, +4), we switch it to adding five points (an impossible event given the rules of the game). If that round subtracted points, we switch it to subtracting four points (also impossible given the rules). We randomly generate 80 cues in this way, leading to impossible cues with final scores including -12, -11, -10, -9, +11, +12, +13, and +14. A total of 310 participants are shown one of these impossible cues before stating their beliefs.

Over and above the fact that all cues are already objectively irrelevant, we make the impossibility of these rule-breaking cues extremely salient to participants (see Appendix A for more details), and therefore a natural supposition is that they should not affect beliefs. Nonetheless, our model suggests that cues depicting impossible outcomes may nonetheless distort beliefs by bringing to mind possible states similar to them. The middle panel of Figure 4 shows that, indeed, positive-but-impossible cues are more similar to positive-and-possible states than negative-but-impossible cues are. Thus, we should expect positive-but-impossible cues to boost beliefs about the probability of positive outcomes. This is exactly what we find, as shown in the right panel of Figure 4. Those shown an impossible-but-positive cue believe that the likelihood of a positive score is 54.6%, compared to only 43.8% for participants shown an impossible-but-negative cue ($p < 0.001$). Thus even explicitly fictional outcomes can shape beliefs, seemingly by shifting which possible outcomes come to mind.

Odd vs Even Cues We have seen that, when cues differ greatly in their representativeness toward a hypothesis, beliefs seem to shift in tandem. The discussion in Section 2.3 suggested that, by this same logic, some beliefs should be relatively impervious to cues. More specifically, when states consistent with a hypothesis are no more similar to each other than they are to those inconsistent with it, no cues can differentially skew the states that come to mind toward or away from that hypothesis. To test this, we asked participants their beliefs about the share of games that end with an even vs odd final score. Intuitively, because odd and even games are "interleaved," being cued with an even outcome will also bring to mind nearby odd outcomes, and *vice versa* for odd cues. Of course, this relies on an intuition about the features that are salient to participants when they assess similarity: if whether an outcome is odd vs even is extremely salient, then odd vs even cues could have substantially different effects on beliefs.

Figure 4: Effect of Impossible Cues on Beliefs about P(Final Score > 0)



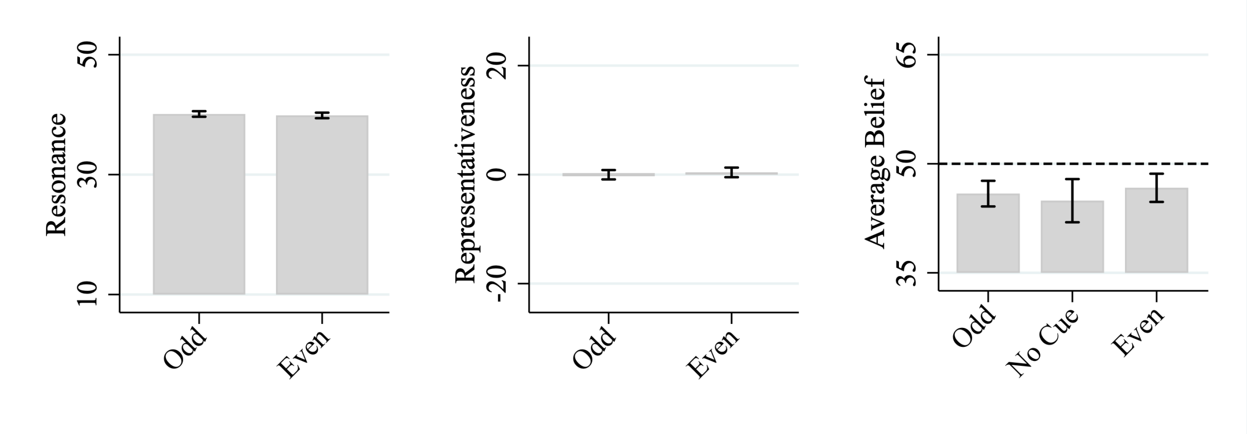
Notes: This figure shows average resonance and representativeness of our baseline "impossible" positive and negative cues (left and middle panels, respectively), each computed from our similarity elicitations. The right panel shows average beliefs across participants shown these cues, as well as the control group, about the probability that the game ends with a positive score.

Our similarity data allow us to check this intuition in a disciplined manner. In particular, the left and middle panel of Figure 5 show that odd and even baseline cues (including both possible and impossible cues) have on average the same resonance ($p = 0.465$) and representativeness with respect to even outcomes ($p = 0.500$). Our model therefore predicts no difference in beliefs about the probability of an even outcome depending on whether participants were shown an even vs odd cue. The right panel of Figure 5 shows that this is indeed what we find. The average belief about the probability of an odd outcome is 45.9% for those shown an odd cue and 46.7% for those shown an even cue (p-value for difference = 0.551). Neither of these averages is significantly different from beliefs in the control group (44.9%, p =0.328 and 0.582 for even and odd, respectively). Thus, not only do cues move beliefs when the similarity data suggests they ought to (positive vs negative cues), they largely fail to move beliefs when the similarity data says they ought not to (even vs odd cues).[9]

Process data: Which outcomes came to mind? In our model, states similar to the cues the agent faces come to mind more readily when she is forming beliefs. To provide suggestive evidence pointing toward this mechanism, we asked participants, on a separate page after they had answered both beliefs questions, to indicate an outcome that "you

[9] We show in Table A.II that in the expanded sample of cues that we use to structurally estimate the model, we do find a small (about 3 p.p.) but significant effect of possible even cues on beliefs about the likelihood of even outcomes. We do not find such an effect for the impossible cues, however, and though insignificant the point estimate is in the opposite direction.

Figure 5: Effect of Odd vs Even Cues on Beliefs about P(Final Score is Even)



Notes: This figure shows average resonance and representativeness of our baseline even and odd cues (left and middle panels, respectively), each computed from our similarity elicitations. The right panel shows average beliefs across participants shown these cues, as well as the control group, about the probability that the game ends with an even score.

felt came to mind for you" while forming their beliefs. Treatment effects on these data strongly track effects on overall beliefs. In particular, participants shown a positive baseline cue are 30.1 p.p. more likely ($p < 0.001$) to indicate that a positive outcome came to mind than those shown a negative baseline cue. In contrast, those shown an even baseline cue are no more likely to report that an even outcome came to mind than those shown an odd cue (difference = -0.3 p.p, $p = 0.884$). See Appendix A for more details on these data.

3.3 Structural estimation

The results above show reduced-form evidence largely in line with the model described in Section 2. But how well quantitatively does our model fit the data? To answer this question, we structurally estimate the model on an expanded set of cues. In addition to the 160 baseline and impossible cues described in the previous section, we collected data on beliefs in response to 200 other cues, chosen partly at random. In particular, for each possible final score (besides those already included in our baseline cues) we randomly selected two cues from the set of possible outcomes with that score. There were two exceptions to this procedure. First, we instead sampled 40 cues each (instead of only two) with final scores of -1 and +1 to have more precise data on cues close to the boundary of our main beliefs question (positive vs negative outcomes). Second, there is only one possible cue each for the most extreme outcomes (final scores of -42 and +28). Because we also wanted more precision at these extremes, participants were more likely to be assigned to these extreme cues than to any other particular cue. Excluding these extreme cues, on average 4 participants were exposed to each cue, whereas 162 and 154 participants were assigned to the extreme negative and positive cues, respectively.

In total, we estimate the model using data from 1,862 participants answering two different questions (about the likelihood of positive vs negative and even vs odd outcomes) facing 360 distinct cues, plus the control group who saw no additional cue. We estimate the parameters of the model therefore to match 722 moments in the data. Our model, in contrast, has only three degrees of freedom, each pertaining to the unobserved default context: its resonance, representativeness with respect to positive vs negative outcomes, and representativeness with respect to even vs odd outcomes.

Similarity Data: Resonance and Representativeness The other inputs to the model—the resonance and representativeness of each of the cues—are pinned down by the similarity judgments that we elicit (i.e., separately from the beliefs data). Panel A of Figure 6 shows the estimated resonance of all cues, broken up by their final score, which show an intuitive pattern. Outcomes closer to "the middle," with final scores near zero, have more resonance. This makes sense given that participants in large part judge similarity by looking at the final score of outcomes (see Table A.I).

In contrast to their resonance, the representativeness of cues is defined relative to a given hypothesis. Thus for each cue, we need to estimate its representativeness both with respect to positive outcomes and with respect to even outcomes, because these are the two beliefs we ask participants about. In Panel B of Figure 6, we see that representativeness with respect to positive outcomes is approximately monotonically increasing in the final score of cues. This result is intuitive: as a cue's final score becomes more positive, the "nearby" states tend to be positive, leading to a higher representativeness. Note that these two patterns imply a trade-off between resonance and representativeness for this question. Toward the middle of the distribution, there are many cues with approximately equal resonance but very different representativeness, suggesting that we should expect substantial differences in beliefs across these cues. But toward the extremes, cues become more or less representative of positive outcomes at the expense of also becoming less resonant. We should therefore expect smaller differences (if any) across beliefs in these ranges, with even a potential for "backfiring" if the reduction in resonance is sufficiently large relative to the increase in representativeness.

Panel C of Figure 6 shows that, unlike with positive vs negative outcomes, cues tend to have approximately zero representativeness with respect to even vs odd outcomes. In fact, only 6% of cues have a representativeness toward even outcomes that is statistically distinguishable (at $p < 0.05$) from zero, indicating that this distribution is quite similar to what we would expect if they all were in fact zero.[10] We should thus largely expect no differences in beliefs about odd vs even outcomes across cues.

Parameters Estimates and Fitted vs Empirical Beliefs We use the generalized method of moments to estimate the model's three free parameters and construct confidence intervals by bootstrapping the data. We estimate that the representativeness of the default toward positive and even outcomes is 0.002 (95% CI = [-0.014, 0.017]) and -0.063 (95% CI = [-0.076, -0.052]), respectively. These estimates, as expected, closely match the average (lack of) bias in the control group for the positive vs negative and even vs odd beliefs. The resonance of the default, which governs how effective cues are in changing beliefs, we estimate at 52.5% (95% CI = [40.0%, 67.6%]), slightly larger than that for our baseline cues (see Figure 3).

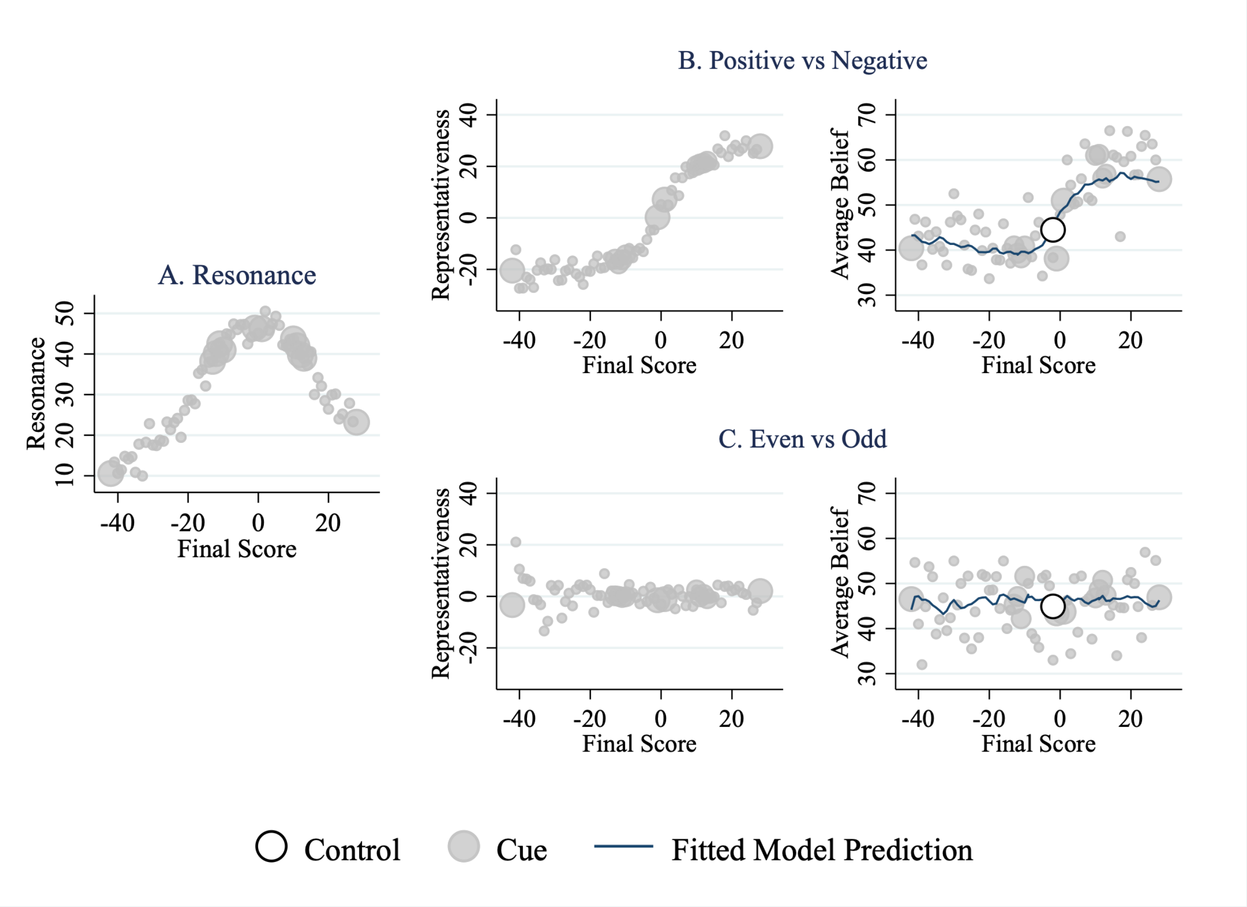
Looking at the full range of cues we tested, the blue curves in Figure 6 show that the model's predictions for beliefs about positive and even outcomes closely track actual average beliefs, as depicted by the gray dots (see Table A.II and the surrounding discussion in Appendix A for a formal regression analysis). In particular, we see in Panel B that for intermediate cues the model expects large increases in beliefs about the likelihood of positive outcomes as the final score increases. This is exactly what we find in the beliefs data: average beliefs increase from 39.8% for baseline negative cues to 51.0% for cues with a final score of +1 ($p < 0.001$, compared to baseline negative) to 58.9% for the baseline positive cues ($p = 0.003$, compared to +1 cues, and $p < 0.001$ compared to baseline negative cues).[11] But, as cues get more extreme, the model expects a tradeoff between representativeness and resonance. In particular, at extreme values, it expects no larger (and, indeed, slightly smaller) effects of cues, despite the more extreme representativeness, due to their much lower resonance. This is also exactly what we find. The most extreme cues (final scores of -42 and +28) have directionally (though not significantly) less of an effect than our baseline cues, and in both cases we can reject that the difference in beliefs is as large as the difference in representativeness ($p = 0.024$ for negative cues and $p < 0.001$ for positive cues).[12]

[10] For comparison, 86% of cues have a representativeness toward positive outcomes that is statistically significantly different from zero. That said, we only barely fail to reject ($p = 0.053$) that the representativeness of all cues with respect to even outcomes are jointly zero.

[11] Note that average beliefs among cues with a final score of -1 are surprisingly low given the model estimates: 38.1% compared to 39.8% in the baseline negative cues ($p = 0.452$).

[12] A different way of testing whether these extreme cues are less effective than their representativeness alone would predict is to compute the ratio between the difference in average beliefs for positive vs negative cues and the difference in representativeness for positive vs negative cues. This ratio is 0.512 for baseline cues (beliefs differ by 19.1 p.p. while representativeness differs by 37.3 p.p), while it is only 0.318 for extreme cues (beliefs differ by 15.3 p.p. while representativeness differs by 48.3 p.p.), and this difference

Figure 6: Model Fit: Beliefs about P(Final Score > 0)



Notes: This figure shows average resonance of all cues (Panel A), their representativeness with respect to positive outcomes (left chart of Panel B), their representativeness with respect to even outcomes (left chart of Panel C), average beliefs about the probability of positive outcomes in response to each cue (right chart of Panel B), and average beliefs about the probability of even outcomes in response to each cue (right chart of Panel C). The curves in the right charts of Panels B and C show the fitted model predictions from our structural estimation. For visual simplicity, we show averages across cues with the same final score, with the size of dots corresponding to the sample sizes. This figure drops a single outlier (cues with a final score of 25) from Panel B, where the average belief (among six participants) was 75.8%.

Completeness and Restrictiveness of the Model How "successful" is the model in explaining the data? Fudenberg et al. (2022) suggest that researchers can answer this question by computing the "completeness" and the "restrictiveness" of their model. More precisely, let E_M be the expected prediction loss (according to the true DGP $\pi(D)$) of model $\mathcal{M}$: $E_M \equiv E_\pi[\mathcal{L}(D|\mathcal{M})]$. The completeness of model $\mathcal{M}$ is given by how much $\mathcal{M}$ improves the expected loss relative to a baseline model $\mathcal{M}_0$ (which in our case is the rational benchmark), as a fraction of the improvement given by a "best-performing" model $\mathcal{M}^*$ (which we estimate by random forest). Formally, the completeness of model $\mathcal{M}$, $\kappa(\mathcal{M})$ is given by:

$$\kappa(\mathcal{M}) = \frac{E_{\mathcal{M}_0} - E_{\mathcal{M}}}{E_{\mathcal{M}_0} - E_{\mathcal{M}^*}} \in [0,1] \tag{8}$$

Even a highly complete model, in the sense of equation 8, may be unsatisfying unless it is also restrictive. That is, if a model is sufficiently flexible that it could explain *any* data, it would not be surprising if it provided a complete explanation of any particular dataset. Following Fudenberg et al. (2020), we estimate the restrictiveness of the model by calculating its completeness with respect to synthetic data. In particular, we create a "jumbled" dataset D^o, where we match each cue with the resonance and representativeness estimates of a randomly selected other cue. Let $E^o_{\mathcal{M}}$ be the expected loss (again estimated using 10-fold cross-validation) on the jumbled dataset. Then, our restrictiveness measure becomes:

$$\rho^{\mathcal{M}} \equiv \frac{E^o_{\mathcal{M}}}{E^o_{\mathcal{M}_0}}. \tag{9}$$

In other words, if our model was was able to fit the jumbled synthetic data much better than the default benchmark, we would conclude that our model is not restrictive – any similarity assessments could be used to justify the observed beliefs data.

We set $\mathcal{M}_0$ as giving the correct Bayesian answer (i.e. the correct and constant answer across all treatments). We estimate the expected loss of each model $\mathcal{M}_0, \mathcal{M}, \mathcal{M}^*$ based on a 10-fold cross-validation. We estimate that our model is 77.3% complete (95% CI = [70.1%, 98.2%]) and 97.6% restrictive (CI = [97.4%, 98.2%]). That is, it achieves 77.3% of the additional explanatory power (over the Bayesian benchmark) of the machine learning estimator despite many fewer degrees of freedom. For comparison, Fudenberg et al. (2022) study the completeness of different classic behavioral theories: they estimate that gambler's/hot-hand fallacy is 7-10% complete in explaining subjects' generation of "random" outcomes, that Poisson cognitive hierarchy models (PCHMs) are 68-97% complete in explaining initial play in games, and that cumulative prospect theory (CPT) is 94%

in ratios is statistically significant ($p = 0.025$).

complete in predicting certainty equivalents. Our model thus achieves completeness (in our novel domain) within the range of these more established theories. But this high level of completeness is not a result of the model being capable of explaining *any* data: the restrictiveness measure indicates that the model only improves on the Bayesian benchmark by 3.4% in our synthetic data. For comparison, Fudenberg et al. (2020) find that CPT is only 27% restrictive, while PCHMs are about 97% restrictive.

3.4 Multiple Cues

What effect does providing multiple additional cues have on beliefs? Recall that participants were shown either one, five, or ten baseline cues (either all positive or all negative). These cues were chosen randomly from the set of positive or negative baseline cues, and thus each cue a participant sees has the same representativeness in expectation. We can thus test whether belief distortions continue to move toward this representativeness as the number of cues increases, and the rate at which they do so. Recall from the discussion in Section 2.3 (and equation 5) that we should expect additional cues to further shift belief distortions toward (but not past) the representativeness of cues.

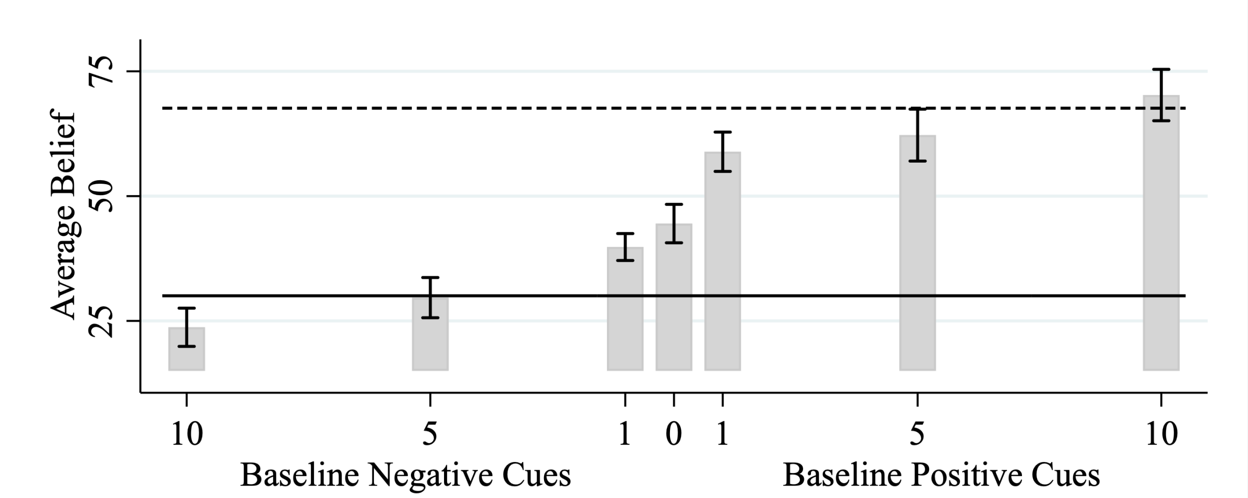
Figure 7 shows average beliefs across these treatments. Focusing first on the positive baseline cues, we see that average beliefs increase slightly from 58.9% for a single cue to 62.2% for five cues, though this differences is not statistically significant ($p = 0.316$). With ten cues, they further increase to 70.2% (different from a single cue at $p < 0.001$ and from five cues at $p = 0.032$), close to the implied maximal distortion of 67.9% (p–value for difference = 0.372). For negative cues, moving from one to five negative cues reduces beliefs from 39.8% to 29.7% ($p < 0.001$), approximately reaching the maximal representativeness-implied distortion, which for the baseline negative cues is 30.7%. Moving to ten cues reduces beliefs by a further 5.9 percentage points ($p = 0.037$) and therefore has a slightly larger effect than our model predicts.

3.5 Discussion: Ruling out Other Explanations

We interpret our results as showing that cues depicting particular states shift beliefs by making it easier to imagine similar outcomes happening. Here, we briefly discuss how our design and results rule out several alternative interpretations of our results.

Learning Many features of the design and results appear inconsistent with a standard learning story. First, there is no objective uncertainty to begin with, as the right answers to all questions are logically pinned down by the description of the data-generating pro-

Figure 7: The Effect of Multiple Cues on Beliefs about P(Final Score > 0)



Notes: This figure shows the maximal representativeness-implied distortion from our baseline positive and negative cues (dashed and solid lines, respectively), and average beliefs about the probability of positive outcomes in response to zero, one, five, and ten negative and positive baseline cues.

cess. Second, our use of a familiar randomization device with a known distribution of outcomes—a fair six-sided die—ensures that participants do not wrongly think the likelihood of a particular state is uncertain. Third, participants are told (and confirm their understanding of the fact that) cues are not informative. Fourth, we restrict our analysis to the (large majority of) participants who display excellent comprehension of all instructions. Fifth, cues do not simply affect beliefs in a "naive" way: e.g., increasing when they depict outcomes consistent with a hypothesis and not otherwise. As we saw, only some beliefs (positive vs negative) and not others (even vs odd) are susceptible to cueing, and the magnitude of these effects are approximately as predicted by the model. Finally, even cues inconsistent with the hypothesis (impossible cues) shift beliefs when they are similar to states consistent with it.

Noise and inattention Next, our results are also inconsistent with being driven by noise or inattention. In addition to the fact that we focus on participants who score perfectly on all comprehension questions (and thus to that extent cannot be described as inattentive), typically noisy or inattentive data are thought to attenuate the responsiveness of decisions relative to an objective benchmark (e.g., Rigotti et al. 2023; Enke, Graeber, et al. 2024). In contrast, we study (and predict) *over*reaction to treatment variation, because the objective benchmark is that beliefs should be insensitive to cueing.

Experimenter demand Finally, while experimenter demand effects—participants anticipating hypotheses and reacting to treatment variation simply out of altruism toward the researchers—are not a salient concern in economics lab experiments (De Quidt et al. 2018, Winichakul et al. 2024), several facts additionally speak against this force driving our results. First, we explicitly tell participants that cues should not affect their beliefs, so if anything experimenter demand may push against finding any effects of cues. This is particularly true in the case of impossible cues, where we stress that the cues depict outcomes that are not possible given the rules of the game. Second, if experimenter demand were driving our results, we might expect all cues (e.g., even vs odd, positive vs negative) to have similarly sized effects. Instead, however, only some cues are effective at changing beliefs: in particular, these tend to be the ones the theory predicts should matter more.

4 Economic applications

Having provided experimental evidence on the model's main predictions, we now turn to investigating its economic implications. We first study exogenous cues. In particular, we focus on the case of intertermporal choice, where an agent's dynamic consumption decisions are distorted by her past or current experiences which act as cues. We show that our model naturally predicts experience effects, projection bias, and present bias, each of whose strength depends on the similarity between cues and future states. We then turn to endogenous cues, where a persuader jointly optimizes a product / service alongside a cue to boost the receiver's willingness to pay. When applied to financial persuasion, we document a novel tradeoff between risk and *persuade-ability*: cue-based persuasion most strongly benefits portfolios with risky / state-dependent returns.

4.1 Exogenous cues: application to intertemporal choice

Setup We consider the following simplified intertemporal setting. Let $\omega_t, \omega_{t+1}, ...$ follow an exogenous Markov process with *known* transition probability $\pi(\omega_{t+1}|\omega_t)$. At time t, the agent chooses a consumption plan $\vec{c_t} = (c_t, c_{t+1}, ...)$, subject to an intertemporal budget constraint $\sum_{h=0}^{\infty} \frac{c_{t+h}}{R^h} \leq W_t$, where R is the exogenous rate of return.The agent's objective expected utility is given by:

$$U_t(\vec{c_t}|\omega_t) = u(c_t|\omega_t) + \sum_{h=1}^{\infty} \beta^h \cdot \left[\sum_{\omega_{t+h} \in \Omega} \pi(\omega_{t+h}|\omega_t) u(c_{t+h}|\omega_{t+h}) \right], \tag{10}$$

where β is the discount factor, and the agent's marginal utility of consumption is potentially state-dependent. Equation 10 implies a standard Euler equation:

$$1 = \beta R \cdot E_t\left[\frac{u'(c_{t+1}|\omega_{t+1})}{u'(c_t|\omega_t)}\right] = \beta R \cdot E_t[M(\omega_{t+1})], \tag{11}$$

where $M(\omega_{t+1}) \equiv \frac{u'(c_{t+1}|\omega_{t+1})}{u'(c_t|\omega_t)}$ is the marginal utility of consumption in state ω_{t+1} relative to the current marginal utility of consumption.

Context and cues Note that rationally solving the intertemporal objective in equation 10 would require the agent to integrate utility from all possible future states over all time horizons. We assume for our agent, states instead come to mind in proportion to their similarity to her context C_t, as in equation 2. We further assume that the cues in context C_t are a subset of the agent's current and past experienced state.[13]

$$C_t \subset \{\omega_{t-h}\}_{h\geq 0}. \tag{12}$$

These assumptions imply that rather than optimizing her true objective function, given by equation 10, the agent instead chooses c_t to effectively optimize:

$$U_t^s(\vec{c_t}|\omega_t, C_t) \propto S(\omega_t, C_t) \cdot u(c_t|\omega_t) + \sum_{h=1}^{\infty} \beta^h \left[\sum_{\omega_{t+h}\in\Omega} \pi(\omega_{t+h}|\omega_t) S(\omega_{t+h}, C_t) u(c_{t+h}|\omega_{t+h})\right]. \tag{13}$$

Proposition 2 characterizes how the context distorts intertemporal consumption:

Proposition 2. *Let $\tilde{S}(\omega_{t+1}) \equiv \frac{S(\omega_{t+1}, C_t)}{S(\omega_t, C_t)}$ be the similarity of ω_{t+1} to C_t, normalized by similarity of ω_t to C_t. Equation 13 implies the following modified Euler equation:*

$$\beta R \cdot E_t[M(\omega_{t+1})] = \left(E_t[\tilde{S}(\omega_{t+1})]\right)^{-1} - \beta R \cdot \mathcal{D}^M(C_t), \tag{14}$$

where $\mathcal{D}^M(C_t) = \frac{Cov_t[M(\omega_{t+1}), S(\omega_{t+1}, C_t)]}{E_t[S(\omega_{t+1}, C_t)]}$ is the representativeness of C_t with respect to future marginal utility.

Note that in the absence of similarity distortions, $\mathcal{D}^{M(\omega_{t+1})} = 0$ and $\tilde{S}(\omega_{t+1}) = 1$, which recovers the normative benchmark (equation 11). Under similarity-based distortions,

[13] One can nest many forms of existing models by extending our model to allow for exogenous or endogenous weights on each past experienced states. For example, by setting $w(\omega_{t-h}) = \delta^{-h}$, our model is equivalent to models of fading memory and experience effects (Malmendier & Nagel, 2016). One can also microfound $w(\omega_{t+h})$ through associative memory models (Kahana, 2012; Bordalo et al., 2023; Bordalo, Burro, et al., 2024), where $w(\omega_{t-h}) \propto S(\omega_{t-h}, \omega_t) \cdot \theta(\omega_{t-h})$, where $\theta(\omega_{t-h})$ is the weight of memory ω_{t-h}.

Proposition 2 implies that there is excessively low current consumption if the current context is representative of high marginal-utility states in the future (high $\mathcal{D}^{M(\omega_{t+1})}(C_t)$). Conversely, there is excessively high current consumption if future states are less associated with C_t (low $E_t[\tilde{S}(\omega_{t+1})]$). We discuss each channel separately.

Experience and projection effects The term $\mathcal{D}^M(C_t)$ captures whether the current context is disproportionately associated with high or low marginal utility future states. One can further decompose this term into the current state ω_t and past experiences:

$$\mathcal{D}^M(C_t) = \frac{\mathcal{W}(\omega_t)}{\mathcal{W}(\omega_t) + \mathcal{W}_{past}} \mathcal{D}^M(\omega_t) + \frac{\mathcal{W}_{past}}{\mathcal{W}(\omega_t) + \mathcal{W}_{past}} \mathcal{D}^M(C_t^{past}), \tag{15}$$

where $C_t^{past} = C_t - \{\omega_t\}$ and $\mathcal{W}_{past} = \sum_{h \geq 1} \mathcal{W}(\omega_{t-h})$. Equation 15 shows how cueing and similarity can provide a unifying framework for two well-established findings in behavioral economics. First, it predicts experience effects (Malmendier, 2021): $\mathcal{D}^M(C_t^{past})$ increases when individuals have experienced past states where the marginal utility of consumption was high (e.g., economic downturns). Second, it predicts projection bias (Loewenstein et al., 2003): $\mathcal{D}^M(\omega_t)$ increases during transient adverse states, illustrating how individuals implicitly over-extrapolate current conditions because it is easier to imagine states similar to the present.

Beyond offering a shared framework for these effects, Equation 15 suggests that the degree to which they shape consumption decisions depends on similarity: experience effects and projection bias will be more pronounced when past or current circumstances especially resemble certain future possible outcomes. Using similarity data to measure these associations may therefore provide a way of generating predictions about the domain specificity of experience effects and projection bias. For related ideas, see Bordalo, Burro, et al. (2024), whose model of memory retrieval and simulation complements our framework.

Example: inconsistent risk attitudes We explore the implications of equation 15 with the following simple example. Let Ω consist of three disjoint subsets, $\Omega = H_{normal} \cup H_{risk,1} \cup H_{risk,2}$: with probability π_n, π_r, and π_r respectively. While the marginal utility of consumption is given by $M(\omega) = C_{t+1}^{-\gamma}$ for $\omega \in H_n$, both $H_{risk,1}$ and $H_{risk,2}$ are "risky" and imply a higher marginal utility of consumption: $M(\omega) = T \cdot C_{t+1}^{-\gamma}$ for $\omega \in H_{risk,1} \cup H_{risk,2}$, with $T > 1$. Lastly, we assume $S(\omega_{normal}, \omega_{risk,1}) = S_{r_1,n}$ and $S(\omega_{risk,1}, \omega_{risk,2}) = S_{r_1,n_2} \leq 1$, for $\omega_{risk,1} \in H_{risk,1}, \omega_{risk,2} \in H_{risk,2}$, and $\omega_{normal} \in H_{normal}$.

Consider an agent who has just experienced a particular type of risk: $\omega_t \in H_{risk,1}$, with the context consisting of the most recent state: $C_t = \{\omega_t\}$. In this setting, one obtains:

$$\mathcal{D}^M(\{\omega_t\}) \propto 1 + S_{r_1,r_2} - 2S_{r_1,n} \tag{16}$$

As equation 16 illustrates, $\mathcal{D}^M$ is decreasing in $S_{r_1,n}$: the harder it is to simulate the normal state (after experiencing $\omega_{risk,1}$), the higher the $\mathcal{D}^M$ (oversaving), consistent with classic scarred consumption and projection of bad states.

However, $\mathcal{D}^M$ is *increasing* in S_{r_1,r_2}: in the extreme, if S_{r_1,r_2} is sufficiently low, there may be *undersaving* relative to the normative benchmark, even if the agent has just experienced a high marginal utility state. This is because if the two risky states are sufficiently dissimilar, experience of one type of risk *crowds out* the simulation of the other type of risk. Such effects may be exacerbated with complete markets: when one allows for the sale of Arrow-Debreu securities (e.g., insurance contracts that pay off in $H_{risk,1}$ or $H_{risk,2}$), one can easily derive that the agent excessively protects herself against the risk she experienced but under-insures against the other risk if S_{r_1,r_2} is sufficiently low. For example, someone cued to think of the risk of natural disasters (e.g., hearing a friend's home was damaged by a hurricane) will begin to act more risk averse in that domain (e.g, buying flood insurance) but if anything might begin to appear less risk averse in dissimilar domains (e.g., home security systems or renter's insurance).

Cognitive discounting The preceding examples highlight that when an agent imagines the future, she disproportionately simulates states of the world similar to her context. However, equations 13 and 14 also highlight a second dimension: the extent to which future states at different horizons come to mind *at all* (relative to current consumption) also depends on similarity, and in particular on how dissimilar the present is from the future. Formally, the second term of equation 14, $(E_t[\tilde{S}(\omega_{t+1})])^{-1}$, captures how easy it is for the individual to simulate consumption in the future. The harder it is for the agent to think of the future (lower $E_t[\tilde{S}(\omega_{t+1})]$), the higher consumption today is relative to the normative benchmark.

To illustrate this, consider the simple case in which the utility of consumption $u(c_t|\omega_t)$ does not depend on ω_t. Despite being normatively irrelevant, ω_t can nevertheless distort consumption by influencing whether and which future states come to mind. Assume that ω_t follows a stationary AR(1) process: $\omega_t = \rho\omega_{t-1} + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, where we impose as before the similarity function $S(\omega, \omega') = \exp\left(-\frac{\kappa}{2}(\omega - \omega')^2\right)$. Equation 13 then implies Proposition 3.

Proposition 3. *The agent effectively optimizes:* $U_t = u(c_t) + \sum_{h\geq 1} \delta_h u(c_{t+h})$, *where the implicit*

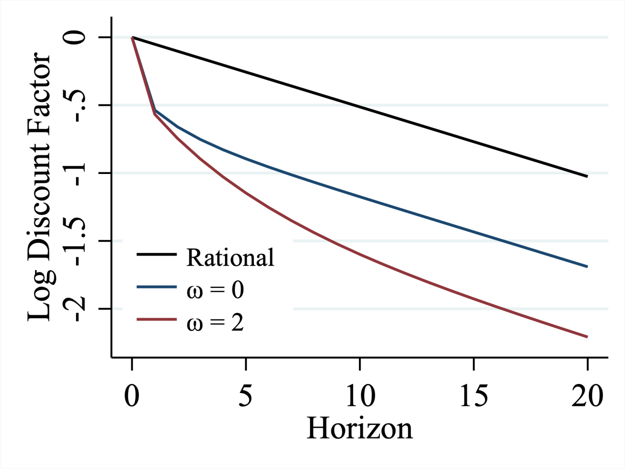


Figure 8: Endogenous cognitive discounting

decision weights δ_h are given by

$$\delta_h = \beta^h \cdot \sqrt{\frac{\tau}{\tau+\kappa}} \cdot \exp\left(-\frac{1}{2}\left(\frac{\tau\cdot\kappa}{\tau+\kappa}\cdot(1-\rho^h)^2\omega^2\right)\right),\ \tau = \frac{1-\rho^2}{\sigma^2(1-\rho^{2(h+1)})} \tag{17}$$

Figure 8 plots the as-if time discount factors δ_h. Relative to the normative benchmark (β^h) plotted in black, our agent first displays more apparent impatience. This arises endogenously due to the fact that future states are decreasingly similar to current state in expectation. Importantly, the *degree* to which the future is discounted is state dependent: there is lower discounting when $\omega = 0$, the median state of the stationary distribution, than when $\omega = 2$ (an unusual state). Our model thus implies the novel prediction that apparent impatience should be more severe when agents are in unusual situations, ones where the likely future is dissimilar to the present. Our agent will, for example, be more frugal at the supermarket than while on vacation or during a novel crisis. The intuition for this result is straightforward: while shopping for groceries it is easy for her to imagine the many future expenses she will have to incur (often in the very same location), whereas these future purchases are less likely to come to mind when her present circumstances are more unique.

Lastly, note that the log discount curves in Figure 8 are convex: the relative discounting of consumption in $t+h$ to $t+h+1$ decreases as $h \mapsto \infty$. This upward curve in the discounting function generates dynamic inconsistency in consumption behavior (Laibson, 1994). We endogenously obtain this result from the simple insight that the far future and the very far future are intuitively equally (dis)similar to the present. Formally, for any stationary process ω_t, one obtains:

$$\lim_{h \mapsto \infty} E[S(C_t, \omega_{t+h})] \mapsto \sum_{\omega \in \Omega} \pi_{stat}(\omega) \cdot S(C_t, \omega), \tag{18}$$

where π_{stat} is the stationary distribution of ω_t. Thus, the relative discounting between $t+h$ and $t+h+1$ decreases to β (the rational benchmark), which generates endogenous hyperbolic discounting.

4.2 Endogenous cues: application to persuasion

While the previous application focused on *exogenous* cues – past and current experiences – we now turn to the case of *endogenous* cues, where a persuader chooses the optimal cue to distort a receiver's beliefs. We assume as before that the receiver is naive about the contextual distortion, and in particular does not take into consideration the strategic motives of the persuader, following other work in behavioral persuasion (Schwartzstein & Sunderam, 2021).

Setup Consider first the following general case. A persuader (who can be viewed as a monopolist firm) produces a single good for a receiver, which delivers utility $\psi(\omega)$ in state ω. We assume that both $\psi(\omega)$ and $\pi(\omega)$ are known by the receiver (conditional on ω coming to mind): the persuader therefore cannot influence receiver through traditional persuasion, such as by providing objectively informative signals about the frequency of states or the quality of the good in them. Instead, it simply chooses a cue $Q \subset \Omega$ that alters the consumer's context when she considers purchasing the good (we assume the cue Q does not overlap with the agent's existing context C). For example, it can run advertisements that highlight particular states, making them (and, through association, other similar states) more likely to come to mind when its potential customers think about purchasing the good.

As given in Proposition 1, the receiver's subjective willingness to pay is given by:

$$E_s[\psi | C \cup Q] = E[\psi] + \frac{\sum_{\omega_c \in C} \mathcal{W}(\omega_c) \cdot \mathcal{D}^{\psi}(\omega_c)}{\sum_{\omega_c \in C} \mathcal{W}(\omega_c)}. \tag{19}$$

We assume the persuader extracts the full surplus from the exchange: it is able to charge the receiver their willingness to pay for the good net of reserve utility ψ_0. The persuader thus solves a joint optimization problem over both the product design (ψ) and the optimal cue (Q), where it maximizes the receiver's subjective willingness to pay net of the cost of producing the good $c(\psi)$:

$$\left(Q^*, \psi^*\right) = \underset{Q,\psi,E_s[\psi|C\cup Q]\geq\psi_0}{\operatorname{argmax}} \quad E_s[\psi|C\cup Q] - c(\psi). \tag{20}$$

Associative cue and product choice Equation 20 implies that the persuader's choice of the underlying product design (ψ) tilts towards goods that are *complementary* with the technology of associative cues. For what kind of goods will there be an effective cue? By the Cauchy-Schwarz inequality, the upper bound on the persuasiveness of a cue is given by the following expression:

$$E_s[\psi|C\cup Q] - E[\psi] \leq \mathcal{D}^{\psi}(C\cup Q) \leq \operatorname{Var}\left(\frac{S(\omega, C\cup Q)}{E[S(\omega, C\cup Q)]}\right)^{1/2} \cdot \operatorname{Var}\Big(\psi(\omega)\Big)^{1/2}. \tag{21}$$

The first term, $\operatorname{Var}\left(\frac{S(\omega,C\cup Q)}{E[S(\omega,C\cup Q)]}\right)^{1/2}$, depends on the associative structure of the state space Ω: there must be variation in how similar states are to each other for cues bring certain states to mind more than others. In particular, as discussed in Section 2.3 and equation 6, associative cues favor goods that deliver high utility in *distinctive* states of the world. The second term, $\operatorname{Var}\Big(\psi(\omega)\Big)^{1/2}$, highlights that associative cues are naturally complementary with products that deliver high utility in specialized states of the world: variation in utility generates scope for an associative cue to inflate the perceived value of the good.

The persuader therefore has an incentive to increase the variation across states in the utility its product delivers, because doing so allows it to cue the receiver with the especially high-utility states. This force leads to more ex post regret by the consumer: she does not find herself in the high-utility states as often as it felt she would when she bought the good. Such an agent might, for example, find herself with a rarely-visited timeshare vacation home, stuck in rush-hour traffic in a high-speed sports car, or frequently paying for product warranties and travel insurance she predictably ends up failing to need.

Application: portfolio choice and underdiversification We can apply these insights to a simplified scenario where a financial advisor extracts fees from an investor by designing a portfolio of assets. In particular, let $\Omega = H_A \cup H_B \cup H_0$, with $\pi(H_A) = \pi(H_B) = p$. H_A and H_B correspond to states of the world where a distinct investment "theme" is realized: for example, there may be an upcoming AI revolution, or a breakthrough in battery technology. States in H_0 correspond to the status quo.[14] Correspondingly, there are two risky

[14] We assume for simplicity these themes cannot co-occur.

assets, also indexed by A and B, with the excess return on each asset given by:

$$R_A = \alpha \cdot I(\omega \in H_A) + \epsilon_A,\, R_B = \alpha \cdot I(\omega \in H_B) + \epsilon_B, \tag{22}$$

with ϵ_A and ϵ_B uncorrelated with variance σ^2. We endow Ω with a similarity structure like that in Section 2.3: there are two features that govern the similarity between states: whether any investment theme has materialized or not, as well as which theme has materialized. This implies $S(\omega_0, \omega_A) = S(\omega_0, \omega_B) = \delta_0 \leq 1$, and $S(\omega_A, \omega_B) = \delta_1 \leq 1$, for $\omega_A \in H_A, \omega_B \in H_B$, and $\omega_0 \in H_0$.

Investors have mean-variance preferences over wealth $u(W) = E[W] - \frac{A}{2} Var[W]$, with access to a riskless saving technology with excess return $R_0 = 0$. They can access risky investments only through their financial advisor. This advisor first provides a cue $Q \subset \Omega$ that highlights some states of the world but does not provide objective information on their likelihood. In particular, the advisor has access to cues of total size $|Q|$, which it can allocate to states in H_A and H_B. We assume for simplicity that the context (absent the persuader's cue) is in the "status quo" state: $C \subset H_0$.[15] Along with the cue, the advisor jointly offers the investor a risky portfolio $x^* = (x^*_A, x^*_B)$, $x^*_A + x^*_B = 1$ at the cost of a fee ϕ^*, which the investor accepts if the subjective expected utility from them exceeds the reserve utility from the riskless asset.[16] Proposition 4 characterizes the optimal persuader action (Q^*, x^*):

Proposition 4. *The persuader adopts a pure cue ($Q \subset H_A$ or $Q \subset H_B$) and offers a portfolio x^* tilted towards the asset that loads on the cued state. The tilt of the portfolio $T(x^*) \equiv \frac{\max(x^*_A, x^*_B)}{\min(x^*_A, x^*_B)}$ is decreasing in the similarity δ_1 between the themes.*

To give an intuition for Proposition 4, Figure 9 plots the maximal fees obtained for a given portfolio (x_A, x_B) as one varies the tilt of the portfolio from $(1, 0)$ to $(0, 1)$. For both the rational benchmark and the persuader-less benchmark, fees are maximized by the fully-diversified portfolio: investors value safety and reward advisors that deliver it. When the advisor can also provide a cue, however, it is the more extreme portfolios (that tilt towards either investment theme) that receive the most boost – as shown by the gap between the blue and gray curves. The intuition is exactly that of Equation 21: there exists more effective cues for the less diversified portfolios. The advisor's optimal

[15] Our persuasion results are unchanged if instead we set C to be the unbiased distribution over Ω, though of course what the investor would choose to do absent an advisor (e.g., whether she would choose not to participate) depends on the choice of default.

[16] Standard mean-variance optimization implies that the advisor offers a portfolio $x^* = \frac{1}{A} Var_s^{-1}[R](E_s[R] - R_0)$ and extracts fees $\phi^* = \frac{1}{A}(E_s[R] - R_0)' Var_s^{-1}[R_i](E_s[R] - R_0)$, where the expectations and variance are taken over the investor's subjective distribution given the context: $\pi_s(\omega) \propto \pi(\omega) \cdot S(\omega, C \cup Q)$.

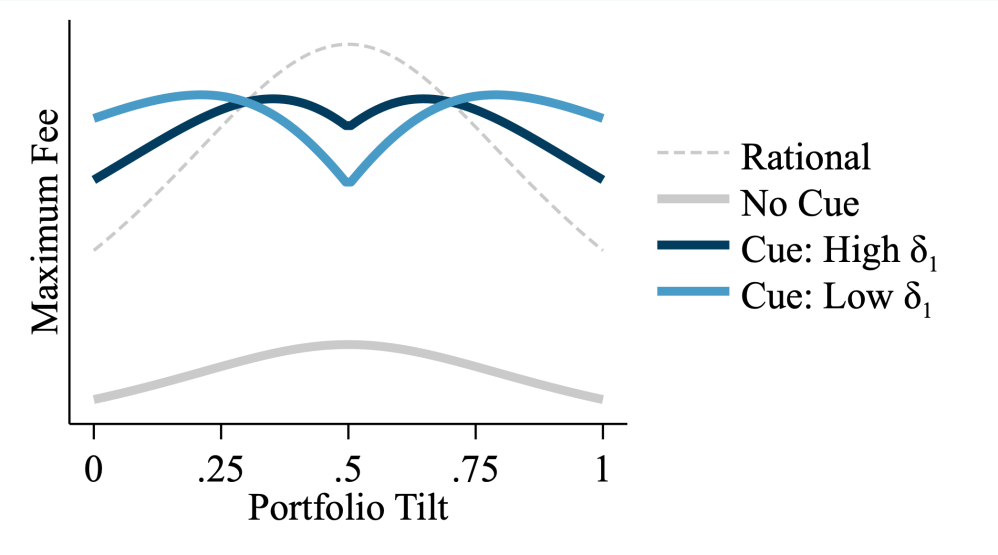


Figure 9: Financial persuasion: portfolio choice

strategy thus tilts the portfolio towards a particular investment theme, optimally trading off persuasiveness with losses from lack of diversification. As we show in Appendix C, this contrasts with the strategy for a purely informative persuader, who offers a fully diversified portfolio while moving beliefs uniformly closer to the truth. Intuitively, while the informative persuader can boost the subjective returns of both assets, a cue-based persuader faces the constraint that cues crowd each other out. In that case, simple (and risky) is best: better to pick a theme and maximally boost its simulation.

Lastly, the optimal portfolio and the degree of diversification depends on the associativity between different investment opportunities: the more similar the themes, the more willing investors are to invest in both opportunities. In particular, under the natural assumption that associativity is positively related to objective correlation—assets that do well in similar circumstances are perceived as more similar—investors will be more persuadable to jointly invest in highly correlated projects, the opposite of what rational diversification would suggest. More generally, this suggests that measuring the associativity between different investment outcomes may be important to understanding financial persuasion and the resulting investment behavior.

Information-cue tradeoff in persuasion Finally, while we have primarily focused on the case of purely uninformative cues, we can also explore how a persuader may combine informative signals with cues. Suppose that the persuader also provides an objective signal I with the likelihood function $\pi(I|\omega) : \Omega \mapsto \mathbb{R}$. We assume that the agent reacts to

I also by sampling states similar to the local context C. That is, given both the context C and information I, the DM's subjective distribution across Ω is given by:

$$\pi(\omega|C,I) \propto S(\omega,C) \cdot \pi(\omega) \cdot \pi(I|\omega). \tag{23}$$

To give an intuition behind Equation 23, the DM is now reacting to the signal in a Bayesian manner – combining the prior and the likelihood – subject to the same cognitive constraint and sampling distortions as before. Equation 23 reduces to our main framework when I is completely uninformative ($\pi(I|\omega) = 1$ for all ω), and also nests Bayesian updating when one removes similarity distortions ($S(\omega,C) = 1$).

Consider the following simple example. Let $\Omega = \mathbb{R}$, with $S(\omega,\omega') = \exp(-\frac{\kappa}{2}(\omega-\omega')^2)$, with prior distribution $\pi(\omega) = \phi_{\mathcal{N}}(\mu_0, \tau_0^{-1})$. The persuader wants to maximize $E_s[\psi(\omega)] \equiv E_s[\omega]$, and has two means at her disposal. First, as before, the persuader can give an uninformative cue Q. Second, the persuader has access to a noisy signal of value I with precision τ ($\pi(I|\omega) = \phi_{\mathcal{N}}(\omega, \tau^{-1})$), which implies a Bayesian posterior:

$$\pi(\omega|I) = \phi_{\mathcal{N}}\left(\frac{\tau_0\mu_0 + \tau I}{\tau_0 + \tau}, (\tau_0+\tau)^{-1}\right) \equiv \phi_{\mathcal{N}}\left(\mu_{post}, \tau_{post}^{-1}\right). \tag{24}$$

Finally, we assume for simplicity that $\mathcal{D}(C) = \mu_{post}$: the subjective mean in absence of the persuader cue is equal to the objective posterior – this is to ensure that any distortion in beliefs is purely due to the persuader-provided cue. These assumptions imply the following closed-form expression for the optimal persuader cue Q^*:

Proposition 5 (Information and cues)**.** *The optimal cue Q^* and agent beliefs are given by:*

$$Q^* = \mu_{post} + \sqrt{\frac{\tau_{post}+\kappa}{\tau_{post}\cdot\kappa}\nu_0} \quad \textit{and} \quad E[\omega|Q^*,I] = \mu_{post} + \sqrt{\frac{\kappa}{\tau_{post}\cdot(\tau_{post}+\kappa)}}\left(\sqrt{\nu_0} - \frac{1}{\sqrt{\nu_0}}\right), \tag{25}$$

where $\nu_0 > 1$ is the unique solution to $1 + \frac{1}{\mathcal{W}(C)}\exp\left(-\frac{1}{2}\nu_0\right) = \nu_0$.

Proposition 5 implies that beliefs $E[\omega|Q^*,I]$ are decreasing in τ_{post} – in particular, for any signal I with precision τ, there exists an alternative signal $I' > I$ with lower precision $\tau' < \tau$ that the persuader prefers, despite it resulting in a *lower* objective posterior mean. This highlights the potential tradeoff faced by a persuader who has access to both information and cues. On the one hand, a positive signal I of high precision will move objective beliefs strongly in the persuader's favored direction. On the other hand, as highlighted by Equation 25, a signal of higher precision limits the scope for persuasion through cues: the more precise the beliefs, the lower the resonance of cues that stray from the objective posterior. Between two signals that yield the same Bayesian posterior mean, the persuader may prefer the one that is noisier, which allows him to further boost beliefs with an effective cue. Our results thus suggest another mechanism for why there may be greater overreaction to weak or extreme signals (Kwon & Tang, 2025; Afrouzi et al., 2023; Bohren et al., 2024; Augenblick, Lazarus, & Thaler, 2024): these are signals highly complementary with an associative cue.

5 Conclusion

In this paper, we offer a theory of beliefs from cues. We show that it provides a parsimonious and structured way of predicting the effect cues have on beliefs. These predictions are largely confirmed in a controlled experimental setting, both concerning the cues our model expects (from independent similarity measures) to matter and concerning beliefs the model expects to be impervious to cueing. Finally, we show that our model can be readily embedded into otherwise standard economic frameworks, and doing so both helps to organize existing behavioral phenomena (experience effects, projection bias, present bias, under-diversification) as well as shed light on how similarity can moderate these effects, generating novel predictions.

Two directions for future work appear especially promising. First, in this paper we focus on an exogenously given and fixed associative structure between states. This is obviously an oversimplification, as the perceived similarity between objects is a function of the features agents happen to be attending to (Tversky 1977). Introducing endogenous attention, and therefore similarity, would enrich the model: cues might not only shape which states initially come to mind, but also which features agents attend to while evaluating their options, which in turn may shape the associations between different states (see Bordalo, Gennaioli, et al. 2024 for related ideas). Second, the usefulness of similarity data in predicting treatment effects in the lab raises the possibility of measuring associations in the field. Which states come to mind when considering a financial investment, or which situations one imagines using a product in, will depend on those states' similarity structures. Measuring and deploying similarity judgments to map out these associations and the corresponding effects of cues "in the wild" are open and exciting avenues for future research.

References

Afrouzi, H., Kwon, S. Y., Landier, A., Ma, Y., & Thesmar, D. (2023). Overreaction in expectations: Evidence and theory. *The Quarterly Journal of Economics*, *138*(3), 1713–1764.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124-150. doi: 10.1037/0033-295X.95.1.124

Augenblick, N., Backus, M., Little, A., & Moore, D. (2024). *Model uncertainty, disagreement, and over-precision: Theory and evidence.* (Working paper)

Augenblick, N., Jack, B. K., Kaur, S., Masiye, F., & Swanson, N. (2023). Retrieval failures and consumption smoothing: A field experiment on seasonal poverty. *Available at Semantic Scholar CorpusID, 263850623*.

Augenblick, N., Lazarus, E., & Thaler, M. (2024). Overinference from weak signals and underinference from strong signals. *Working paper*. Retrieved from http://arxiv.org/abs/2109.09871

Ba, C., Bohren, J. A., & Imas, A. (2023). *Over-and underreaction to information.* Retrieved from https://ssrn.com/abstract=4274617

Barberis, N. C., & Jin, L. J. (2023). *Model-free and model-based learning as joint drivers of investor behavior* (Tech. Rep.). National Bureau of Economic Research.

Barron, K., & Fries, T. (2023). Narrative persuasion. *CESifo Working Paper no. 10206*. Retrieved from www.RePEc.org

Becker, G. S., & Mulligan, C. B. (1997). The endogenous determination of time preference. *The Quarterly Journal of Economics*. Retrieved from https://academic.oup.com/qje/article/112/3/729/1926899

Becker, G. S., & Murphy, K. M. (1993). A simple theory of advertising as a good or bad. *The Quarterly Journal of Economics*, *108*(4), 941–964.

Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, *75*, 450-457. Retrieved from http://dx.doi.org/10.1016/j.neuropsychologia.2015.06.034 doi: 10.1016/j.neuropsychologia.2015.06.034

Bohren, J. A., Hascher, J., Imas, A., Ungeheuer, M., & Weber, M. (2024). *A cognitive foundation for perceiving uncertainty.* Retrieved from https://ssrn.com/abstract=4706147

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., & Shleifer, A. (2024). Imagining the future: Memory, simulation, and beliefs about covid. *Review of Economic Studies*.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., & Shleifer, A. (2023). Memory and probability. *The Quarterly Journal of Economics*, *138*(1), 265–311.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., & Shleifer, A. (2025). How people use statistics. *The Review of Economic Studies*.

Bordalo, P., Gennaioli, N., Lanzani, G., & Shleifer, A. (2024). A cognitive theory of reasoning and choice.

Caplin, A., Dean, M., & Leahy, J. (2019). Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, *86*(3), 1061–1094.

Charles, C. (2022a). *Memory and trading.*

Charles, C. (2022b). *Memory moves markets.*

Charles, C., & Kendall, C. (2024). Causal narratives. *Working paper*.

Chater, N., Zhu, J. Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020, 10). Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, *29*, 506-512. doi: 10.1177/0963721420954801

Colonnelli, E., Gormsen, N. J., & Mcquade, T. (2024). Selfish corporations. *The Review of Economic Studies*.

Conlon, J. (2024). Attention, Information, and Persausion. *Working paper*.

Conlon, J. (2025). Memory Rehearsal and Belief Biases. *Working paper*.

De Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, *108*(11), 3266–3302.

Dougherty, M. R., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, *70*, 135-148. doi: 10.1006/obhd.1997.2700

Eliaz, K., & Spiegler, R. (2020). A model of competing narratives. *American Economic Review*, *110*, 3786-3816. doi: 10.2307/26966480

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

Enke, B., Graeber, T., Oprea, R., & Yang, J. (2024). *Behavioral attenuation* (Tech. Rep.). National Bureau of Economic Research.

Enke, B., Schwerter, F., & Zimmermann, F. (2024). Associative memory, beliefs and market interactions. *Journal of Financial Economics*, *157*, 103853.

Esponda, I., & Vespa, E. (2014). Hypothetical thinking and information extraction in the laboratory. *American Economic Journal: Microeconomics*, *6*, 180-202. doi: 10.1257/mic.6.4.180

Evers, E. R. K., Imas, A., & Kang, C. (2021). On the role of similarity in mental accounting and hedonic editing. *Psychological Review*.

Fudenberg, D., Gao, W., & Liang, A. (2020). *Quantifying the restrictiveness of theories* (Tech. Rep.). Working Paper.

Fudenberg, D., Gao, W., & Liang, A. (2023). How flexible is that functional form? quantifying the restrictiveness of theories. *Review of Economics and Statistics*, 1–50.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, *130*(4), 956–990.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of behavioral economics: Applications and foundations 1* (Vol. 2, pp. 261–343). Elsevier.

Gabaix, X., & Laibson, D. (2022). Myopia and discounting. *NBER Working Paper 23254*. Retrieved from http://www.nber.org/papers/w23254

Gershman, S. J., Zhou, J., & Kommers, C. (2017, 12). Imaginative reinforcement learning: Computational principles and neural mechanisms. *Journal of Cognitive Neuroscience*, *29*, 2103-2113. doi: 10.1162/jocn_a_01170

Gilbert, D. T., & Wilson, T. D. (2007, 9). Prospection: Experiencing the future. *Science*, *317*, 1351-1354. doi: 10.1126/science.1140734

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics*. Retrieved from https://academic.oup.com/qje/article/110/3/605/1859208

Graeber, T., Roth, C., & Zimmermann, F. (2023). *Stories, statistics, and memory.*

Haaland, I., Roth, C., & Wohlfart, J. (2023, 3). Designing information provision experiments. *Journal of Economic Literature*, *61*, 3-40. doi: 10.1257/jel.20211658

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Kahneman, D., & Tversky, A. (1981). The simulation heuristic. *Working paper.1*.

Kwon, S. Y., & Tang, J. (2025). Extreme categories and overreaction to news. *Review of Economic Studies*.

Laibson, D. (1994). *Hyperbolic discounting and consumption* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.

Link, S., Peichl, A., Roth, C., & Wohlfart, J. (2023). Attention to the macroeconomy. *ECONtribute Discussion Paper No. 256*. Retrieved from www.econtribute.de

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *the Quarterly Journal of economics*, 1209–1248.

Malmendier, U. (2021). Exposure, Experience, and Expertise: Why Personal Histories Matter in Economics. *Journal of the European Economic Association*.

Malmendier, U., & Nagel, S. (2011). Depression babies: do macroeconomic experiences affect risk taking? *The quarterly journal of economics*, *126*(1), 373–416.

Malmendier, U., & Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, *131*(1), 53–87.

Malmendier, U., & Wachter, J. A. (2022). Memory of past experiences and economic decisions. *Handbook of Human Memory*.

Martínez-Marquina, A., Niederle, M., & Vespa, E. (2019). Failures in contingent reasoning. *American Economic Review*, *109*, 3437-3474. doi: 10.2307/26789063

Milgrom, P., & Roberts, J. (1986). Price and advertising signals of product quality. *Journal of Political Economy*, *94*, 796-821.

Moser, J. (2019, 7). Hypothetical thinking and the winner's curse: an experimental investigation. *Theory and Decision*, *87*, 17-56. doi: 10.1007/s11238-019-09693-9

Mullainathan, S. (2002). A memory-based model of bounded rationality. *The Quarterly Journal of Economics*. Retrieved from https://academic.oup.com/qje/article/117/3/735/1932979

Mullainathan, S., Schwartzstein, J., & Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly Journal Of Economics*.

Mullally, S. L., & Maguire, E. A. (2014). *Memory, imagination, and predicting the future: A common brain mechanism?* (Vol. 20). SAGE Publications Inc. doi: 10.1177/1073858413495091

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022, 11). Managing self-confidence: Theory and experimental evidence. *Management Science*, *68*, 7793-77817. doi: 10.1287/mnsc.2021.4294

Niederle, M., & Vespa, E. (2023). Cognitive limitations: Failures of contingent thinking. *Annual Review of Economics*, *15*, 307-328. Retrieved from https://doi.org/10.1146/annurev-economics- doi: 10.1146/annurev-economics

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*(1), 54.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annu. Rev. Psychol*, *43*, 25-53. Retrieved from www.annualreviews.org

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American economic review*, *89*(1), 103–124.

Oprea, R. (2024). Decisions under risk are decisions under complexity. *Working paper*.

Rabin, M. (2013). An approach to incorporating psychology into economics. *American Economic Review*, *103*(3), 617–622.

Rigotti, L., Wilson, A., & Gupta, N. (2023). The experimenters' dilemma: Inferential preferences over populations. *Working paper*.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007, 9). *Remembering the past to imagine the future: The prospective brain* (Vol. 8). doi: 10.1038/nrn2213

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012, 11). *The future of memory: Remembering, imagining, and the brain* (Vol. 76). doi: 10.1016/j.neuron.2012.11.001

Schwartzstein, J., & Sunderam, A. (2021). Using models to persuade. *American Economic Review*, *111*, 276-323. doi: 10.1257/aer.20191074

Sims, C. A. (2011). Rational inattention and monetary economics. *Handbook of Monetary Economics*. doi: 10.1016/S0169-7218(11)03004-8

Stantcheva, S. (2022). *How to run surveys: A guide to creating your own identifying variation and revealing the invisible *.* Retrieved from https://amerispeak.norc.org/us/en/amerispeak/about-amerispeak/overview.html

Stern, B. L., & Resnik, A. J. (1991). Information content in television advertising: A replication and extension. *Journal of Advertising Research*, *31*(3), 36–46.

Taubinsky, D., Butera, L., Saccarola, M., & Lian, C. (2024). *Beliefs about the economy are excessively sensitive to household-level shocks: Evidence from linked survey and administrative data *.*

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*, 123-154.

Winichakul, K. P., Lezema, G., Mustafi, P., Lepper, M., Wilson, A., Danz, D., & Vesterlund, L. (2024). The effect of experimenter demand on inference.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601. doi: 10.1146/annurev-economics-102819-040518

A Additional Experimental Details and Results

Here we provide additional details about the experimental design and supplementary results.

A.1 Main supplemental analysis and details

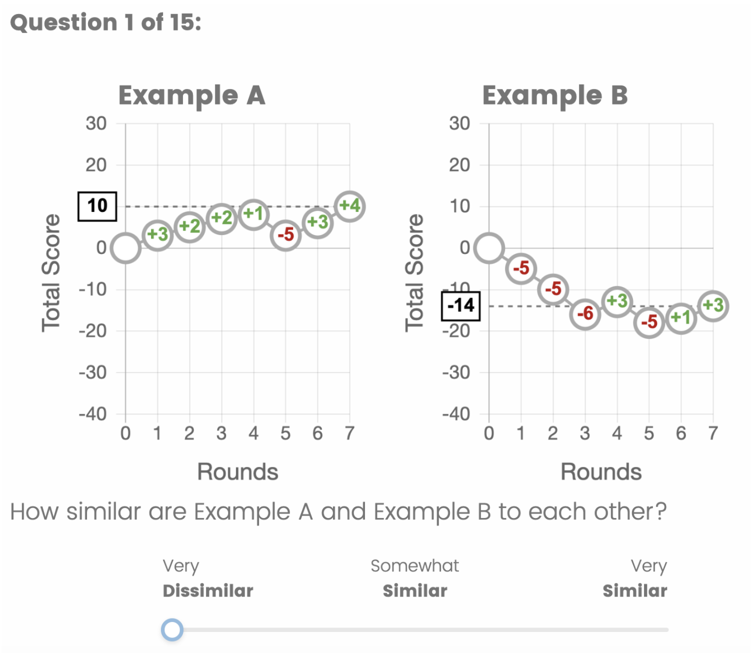
Features of similarity. In Section 3.1, we mentioned that the judged similarity between states in our experiment is to a large extent determined by the distance between their final scores. More precisely, Table A.I shows OLS regressions where the observations are pairs of outcomes that participants rated the similarity of. The dependent variable is how similar a participant judged the two outcomes to be. In column 1 we include only individual fixed effects, while in columns 2 through 7 we add different additional variables, including the absolute value of the difference between the outcomes' final score (column 2), a dummy for whether the final score is positive (column 3), a dummy for whether the final score is even (column 4), their average score across all seven rounds (column 5), their maximum score across all seven rounds (column 6), their minimum score across all seven rounds (column 7). Finally, column 8 includes all these variables in the same specification.

Several facts stand out. First, the distance between final scores has a much higher explanatory power (in terms of R^2) than any other single variable: two paths that end near each other are judged as quite similar. Second, though their additional explanatory in terms of increased R^2 is small compared to only looking at final score, the other variables (other features of the endpoint as well as other aspects of the path such as average levels) also significantly predict similarity judgments. Both (differences in) the positive and even dummies are statistically significant even controlling for final score distance, though the effect size of the positive dummy is more than ten times as large as that for the even dummy. Furthermore, while the partial effects of distances in average, maximum, and minimum scores are not individually significant, we can (marginally) reject that they are all jointly zero ($p = 0.086$). Thus, many features appear to impact similarity in our context, though the largest contributor is the distance between final scores.

Table A.I: Features of Similarity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Abs(FinalScore$_1$ -FinalScore$_2$)		-1.42***						-1.24***
		(0.10)						(0.06)
$\mathbb{1}$(Positive$_1$ $\neq$ Positive$_2$)			-20.41***					-5.49***
			(1.08)					(0.49)
$\mathbb{1}$(Even$_1$ $\neq$ Even$_2$)				-0.65**				-0.46**
				(0.31)				(0.23)
Abs(AverageScore$_1$ -AverageScore$_2$)					-2.16***			0.07
					(0.09)			(0.14)
Abs(MaxScore$_1$ -MaxScore$_2$)						-1.78***		-0.12
						(0.16)		(0.08)
Abs(MinScore$_1$ -MinScore$_2$)							-1.54***	-0.14
							(0.15)	(0.12)
Constant	35.83***	59.97***	46.07***	36.16***	58.00***	51.75***	54.12***	61.71***
	(1.60)	(1.16)	(1.95)	(1.65)	(0.79)	(1.39)	(1.46)	(1.12)
Observations	36,270	36,270	36,270	36,270	36,270	36,270	36,270	36,270
Individuals	2,418	2,418	2,418	2,418	2,418	2,418	2,418	2,418
R^2	0.16	0.53	0.27	0.16	0.46	0.33	0.43	0.54

Notes: This table shows OLS regression estimates. The dependent variable is the similarity rating given by a participant to a pair of outcomes. Each individual rated 15 such pairs. All specifications include individual fixed effects and show standard errors, clustered at the individual and outcome levels, in parentheses. The independent variables in columns 2-8 include the absolute difference in the final score of the two outcomes, whether their final scores' differ in being positive vs not, whether their final scores' differ in being even or not, the absolute difference in their average scores across all seven rounds, the absolute difference in their maximum score across all seven rounds, and the absolute different in the minimum score across all seven rounds. *, **, and *** indicate statistical significance at the $p < 0.10$, 0.05, and 0.01 levels, respectively.



Question 1 of 15:

How similar are Example A and Example B to each other?

Figure A.I: Example of a Similarity Elicitation

Impossible cues. In Section 3.2, we mentioned that the fact that our impossible cues were disallowed by the rules of the game was made extremely salient to participants. The full survey instrument can be accessed at this online document, but here we provide more details about how this was done. The text preceding an impossible cue reads, "You'll notice that **this example game will not follow the rules perfectly**: one of the rounds **adds 5** [or **subtracts 4**] points, which is not one of the possible die rolls." We then include a comprehension question where participants must correctly report during which round a illegal move was made.[17] During the waiting period between revealing the cue and eliciting beliefs, the text reads "In this example game, the final score was X, but remember it **does not perfectly follow the rules**." Then, right above the question eliciting their beliefs, the experiment tells participants to "focus again **only on games that actually follow all the rules**" (all bolded text is as it appears in the experiment).

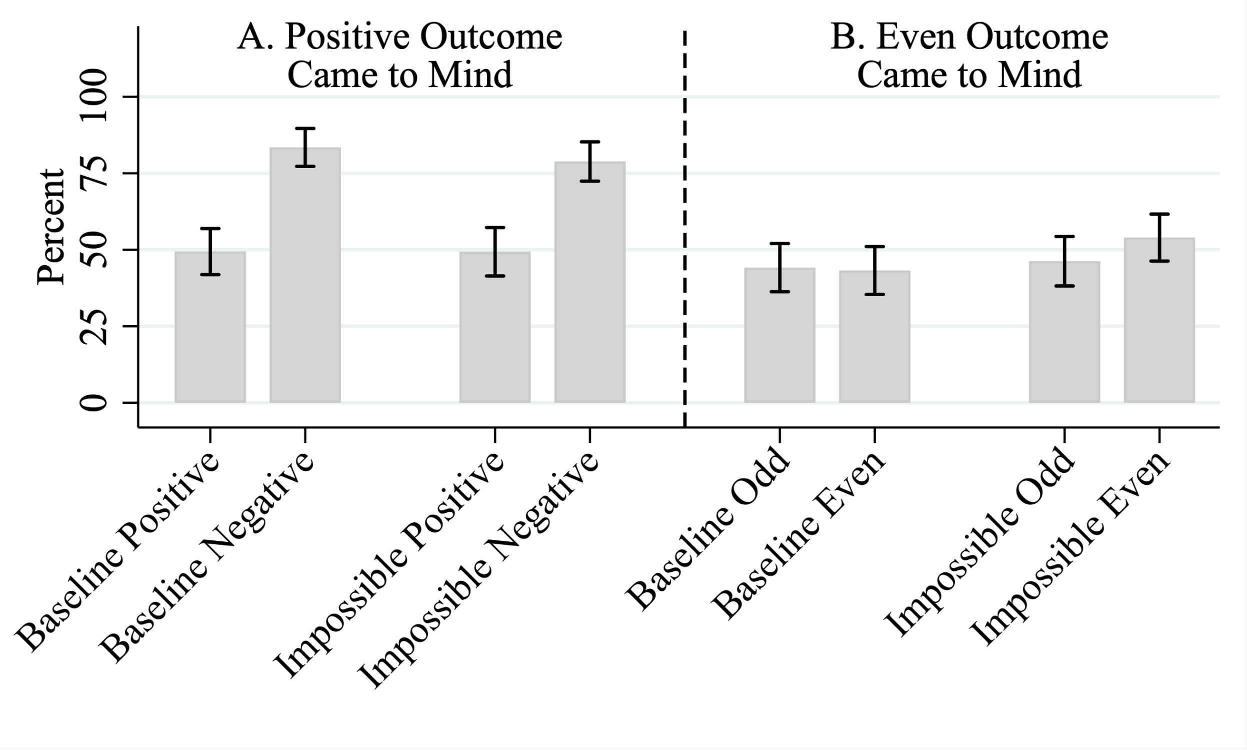
[17] 87% answer this question correctly on the first attempt, and results are unchanged if we focus only on this group.

Process data. In Section 3.2, we described data on which outcomes participants felt “came to mind” when they were forming their beliefs. We elicited these data after participants reported their beliefs. They used an interactive graph to “draw” a game by choosing the number of points added/subtracted in each round. We can thus record whether the outcome they indicate is positive vs negative and is even vs odd, and how this tendency differs by the type of cue they were earlier shown.

Figure A.II shows that these data on which outcomes came to mind mirror our treatment effects on beliefs. Participants who saw a baseline positive cue are 34.0 p.p. more likely to indicate a positive outcome than those who saw a baseline negative cue ($p < 0.001$). For impossible cues, this difference is similar (29.5 p.p., $p < 0.001$), and we cannot reject that these two effects are equal ($p = 0.528$). In contrast, we see no differences in the propensity to indicate an even vs odd outcome depending on cues. Across the baseline and impossible cues, 47.0% of participants indicate that an even outcome came to mind, and we cannot reject that this rate is the same across all treatments ($p = 0.204$). In particular, we cannot reject that more participants indicate an even outcome when shown an even cue, for either the baseline ($p = 0.870$) or impossible ($p = 0.175$) cues.

Note that, while these data are consistent with the mechanism we propose, these process data are not conclusive. In particular, there is no objective benchmark to compare these data to, and so it is unclear how participants “ought” to answer this question. Eliciting which unrealized possibilities come to mind differs in this respect from, say, free-recall measures in related models of memory retrieval (e.g., Bordalo et al. 2023, Conlon 2025), where incentivize-able measures of both beliefs and recall can be independently elicited and compared. Nonetheless, we view our data on which outcomes come to mind as providing suggestive evidence consisent with the model we describe.

Figure A.II: Process Data: Which Outcomes Came to Mind?



Notes: This figure shows the percent of participants who, when asked to indicate an outcome that came to mind when forming beliefs, indicated an outcome with a positive final score (Panel A) and with an even score (Panel B), broken up by the type of baseline cue participants were shown (possible vs impossible, positive vs negative, and even vs odd).

Comparing Model-Predicted and Empirical Beliefs Table A.II shows OLS regressions where the dependent variable is participant's beliefs (about positive outcomes in columns 1-4, even outcomes in columns 5-6, and combining data on both in columns 7-8) and where the independent variables include the fitted model's predictions of those beliefs. We see in column 7 that, combining the two beliefs question we elicit, the model's predictions on average fit the data: for every percentage point increase in beliefs that the model expects, actual beliefs on average increase by 1.00 percentage points. As expected, this fit is largely driven by beliefs about positive outcomes (which the model can explain through differences in representativeness, see column 1) rather than by beliefs about even outcomes (where, given that representativeness toward even outcomes tends to be zero, the model largely expects no differences in beliefs across cues, see column 5).

Furthermore, note that the predictive power of the model appears robust to including proxies for several other natural heuristics that could drive beliefs. For example, columns 2 and 8 show that the model's predictions remain significant even after including dummies for whether the cue shows an outcome that is "consistent" with the hypothesis being asked about (note that the definition of consistency is not especially clear for impossible cues, which are strictly speaking inconsistent with every hypothesis). Similarly, beliefs about positive outcomes do not seem well described by a simple anchoring-and-adjustment story, where participants' beliefs react only to the final score of the cue they are shown (columns 3 and 4).

Table A.II: Predicted vs Empirical Beliefs

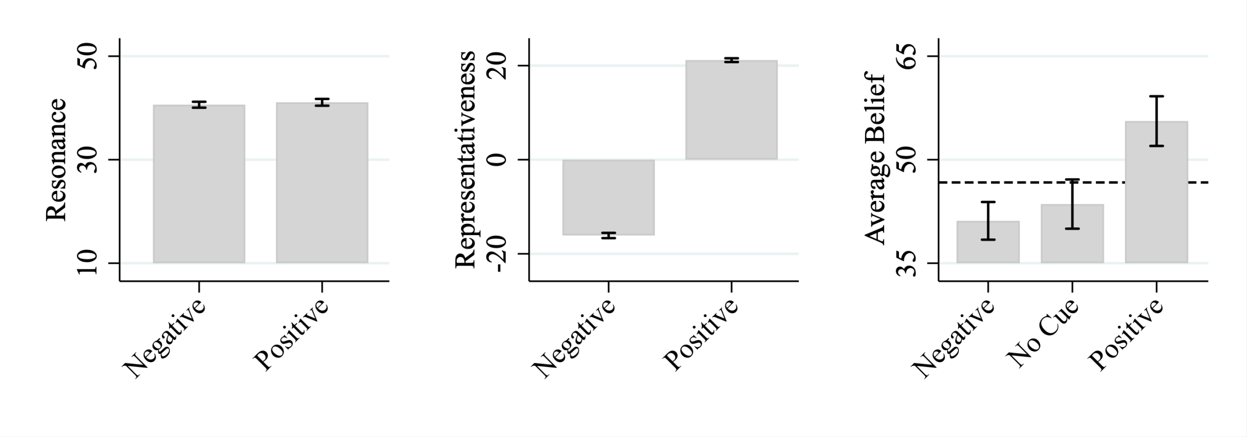
	P(Final Score > 0)				P(Final Score is Even)		Combined	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Model Prediction	1.05***	0.43**	1.05***	0.44**	0.08	0.14	1.00***	0.80***
	(0.07)	(0.17)	(0.12)	(0.18)	(0.20)	(0.23)	(0.07)	(0.07)
Possible × Consistent		10.86***		10.97***		3.18***		4.91***
		(2.29)		(2.51)		(0.84)		(0.75)
Impossible × "Consistent"		3.94		46.48		-1.62		-2.04
		(3.74)		(36.94)		(1.88)		(1.78)
Possible × Final Score			0.05	-0.00				
			(0.03)	(0.03)				
Impossible × Final Score			-0.26*	-1.79				
			(0.13)	(1.54)				
Possible		-4.64***	-1.63	15.60		-1.70		-3.75***
		(1.62)	(1.35)	(17.96)		(1.39)		(1.16)
Control (No Cue)		2.15**	-2.64***	2.22*		1.08		1.58***
		(1.01)	(0.61)	(1.24)		(0.88)		(0.50)
Constant	-2.43	26.70***	-0.52	6.26	42.11***	39.23***	0.13	10.46***
	(3.51)	(7.06)	(6.04)	(19.45)	(9.33)	(10.67)	(3.06)	(3.38)
Observations	1,862	1,862	1,862	1,862	1,862	1,862	3,724	3,724
R^2	0.08	0.10	0.09	0.10	0.00	0.01	0.05	0.06

Notes: This table shows OLS regression estimates. The dependent variable is participants' beliefs about the probability that the game's final score is positive (columns 1-4), even (5-6), and both (in columns 7-8). "Model Prediction" is the fitted structural model's prediction for each belief. "Possible" is an indicator for whether the cue a participant was shown a possible cue. "Consistent" is an indicator for whether the cue was consistent with the hypothesis the participant was considering (i.e., a positive or even outcome). Note that this variable name appears in quotation marks when applied to an impossible cue, as it is unclear whether to consider, for example, a positive-but-impossible outcome consistent with the hypothesis the participant is considering (which concerns only possible outcomes). "Final Score" is the final score of the cue. "Control" is an indicator for whether participants were shown zero cues (the control group). For such participants, we impute values of zero for all other variables in the regression. Only participants who saw zero or one cue are included in these regressions. *, **, and *** indicate statistical significance at the $p < 0.10$, 0.05, and 0.01 levels, respectively.

A.2 Comprehension robustness

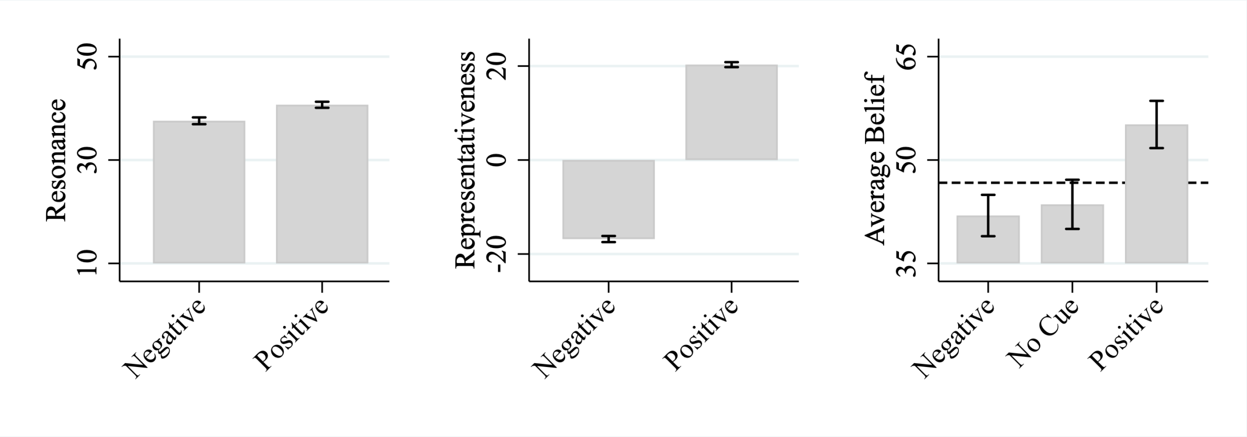
Next, we recreate all the figures from Section 3 but also including the 23% of participants who make any errors on comprehension questions. In each case, the main results are unchanged.

Figure A.III: Effect of Baseline Cues on Beliefs about P(Final Score > 0)



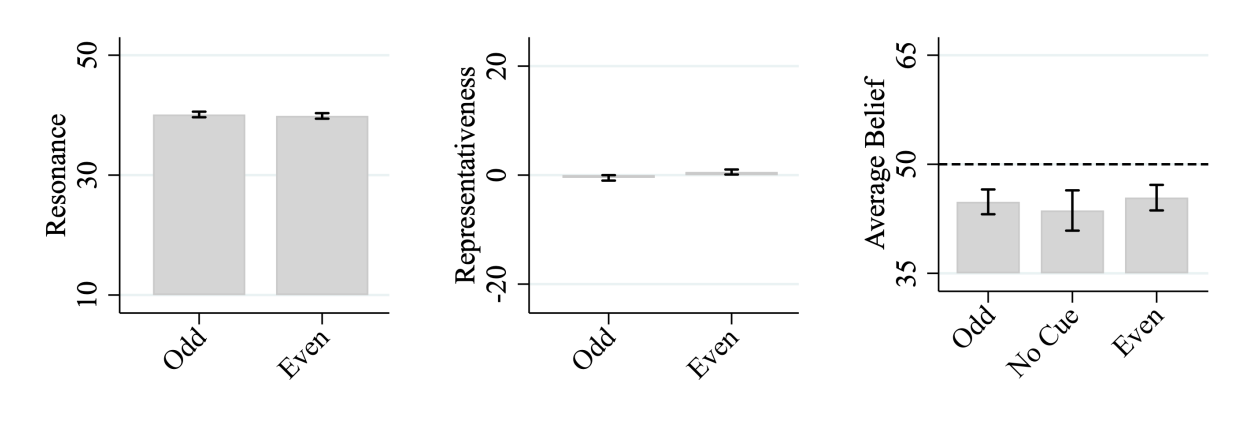
Notes: This figure shows average resonance and representativeness of our baseline positive and negative cues (left and middle panels, respectively), each computed from our similarity elicitations. The right panel shows average beliefs across participants shown these cues, as well as the control group, about the probability that the game ends with a positive score. This figure, unlike the analogous figure in the main text, includes also data from the 23% of participants who made any mistakes on comprehensions questions.

Figure A.IV: Effect of Impossible Cues on Beliefs about P(Final Score > 0)



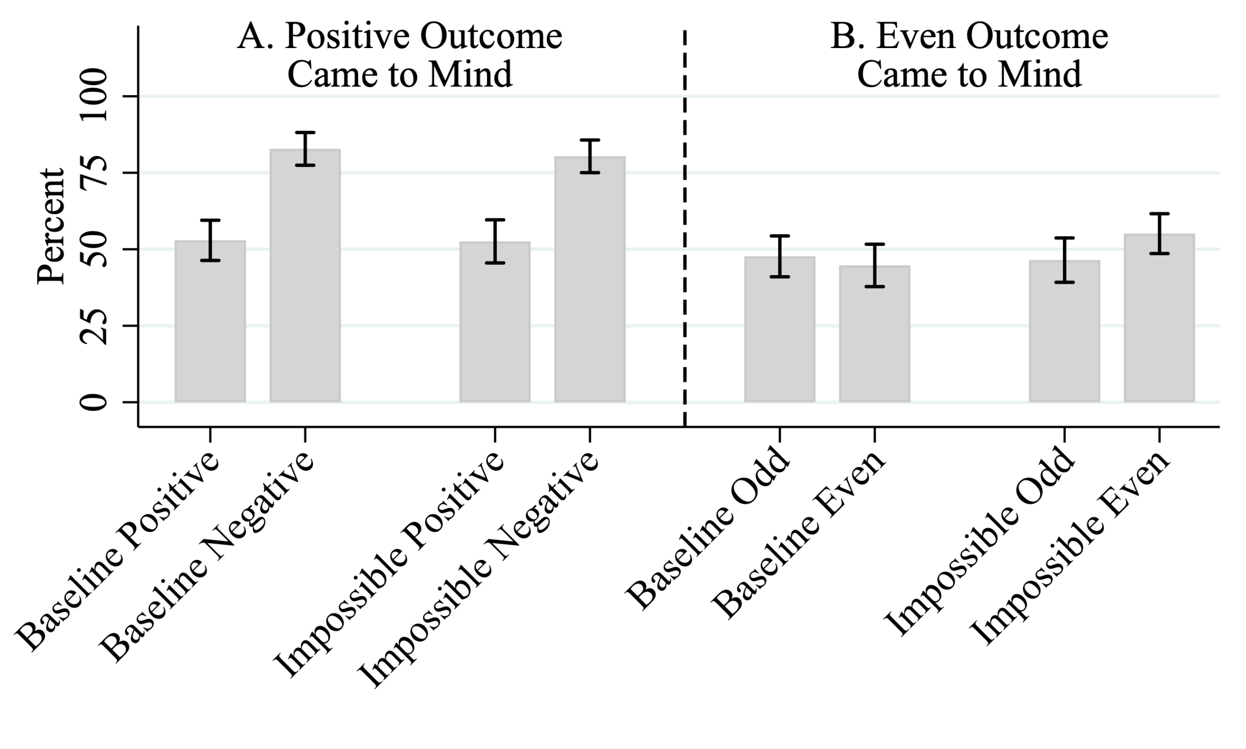
Notes: This figure shows average resonance and representativeness of our baseline “impossible” positive and negative cues (left and middle panels, respectively), each computed from our similarity elicitations. The right panel shows average beliefs across participants shown these cues, as well as the control group, about the probability that the game ends with a positive score. This figure, unlike the analogous figure in the main text, includes also data from the 23% of participants who made any mistakes on comprehensions questions.

Figure A.V: Effect of Odd vs Even Cues on Beliefs about P(Final Score is Even)



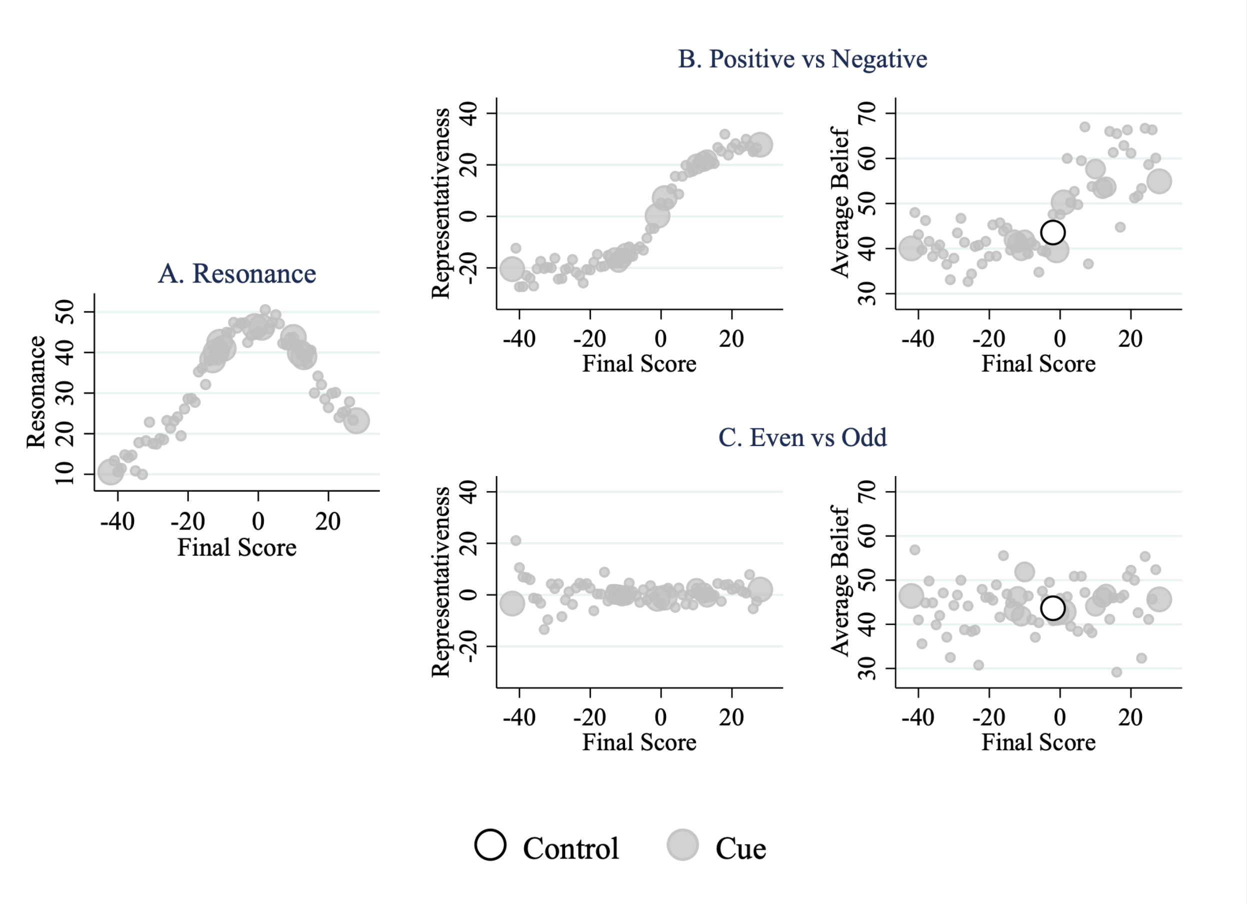
Notes: This figure shows average resonance and representativeness of our baseline even and odd cues (left and middle panels, respectively), each computed from our similarity elicitations. The right panel shows average beliefs across participants shown these cues, as well as the control group, about the probability that the game ends with an even score. This figure, unlike the analogous figure in the main text, includes also data from the 23% of participants who made any mistakes on comprehensions questions.

Figure A.VI: Process Data: Which Outcomes Came to Mind?



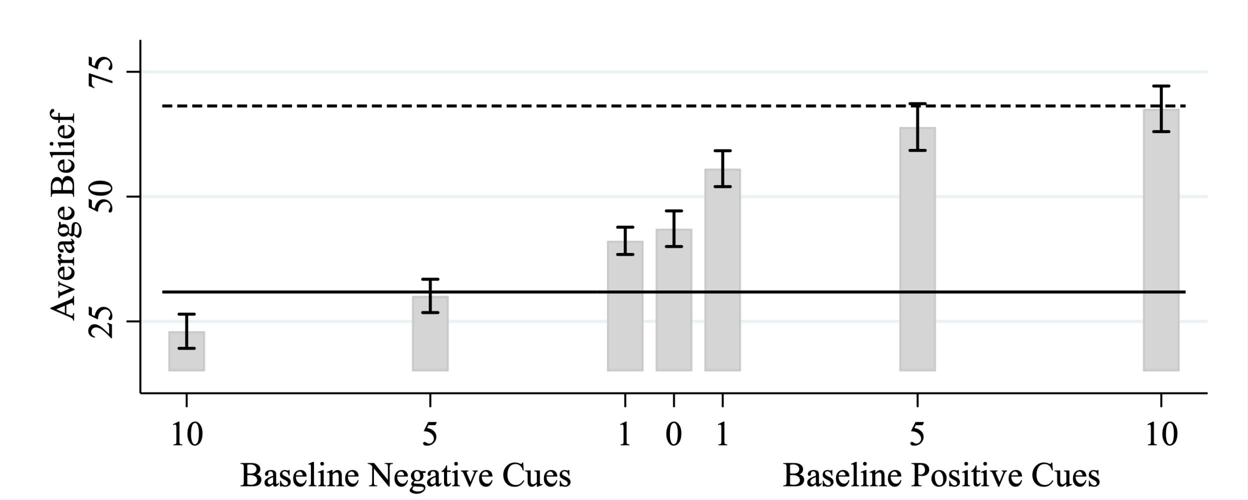
Notes: This figure shows the percent of participants who, when asked to indicate an outcome that came to mind when forming beliefs, indicated an outcome with a positive final score (Panel A) and with an even score (Panel B), broken up by the type of baseline cue participants were shown (possible vs impossible, positive vs negative, and even vs odd). This figure, unlike the analogous figure above, includes also data from the 23% of participants who made any mistakes on comprehensions questions.

Figure A.VII: Model Fit: Beliefs about P(Final Score > 0)



Notes: This figure shows average resonance of all cues (Panel A), their representativeness with respect to positive outcomes (left chart of Panel B), and their representativenss with respect to even outcomes (left chart of Panel C), average beliefs about the probability of positive outcomes in response to each cue (right chart of Panel B), and average beliefs about the probability of even outcomes in response to each cue (right chart of Panel C). The curce in the right charts of Panels B and C show the fitted model predictions from our structural estimation. This figure, unlike the analogous figure in the main text, includes also data from the 23% of participants who made any mistakes on comprehensions questions.

Figure A.VIII: The Effect of Multiple Cues on Beliefs about P(Final Score > 0)



Notes: This figure shows representativeness of our baseline positive and negative cues (dashed and solid lines, respectively), and average beliefs about the probability of positive outcomes in response to zero, one, five, and ten negative and positive baseline cues. This figure, unlike the analogous figure in the main text, includes also data from the 23% of participants who made any mistakes on comprehensions questions.

B Proofs

In all the proofs, we solve assuming the extension of our model to account for weighted subsets. Formally, we define the context $C = \{\omega_c, w(\omega_c)\}$ as a weighted subset of Ω, with similarity between two weighted subsets of Ω, A and B, defined as: $S(A,B) = \frac{1}{|A|\cdot|B|}\sum_{\omega,\omega'\in\Omega} w_A(\omega)w_B(\omega')\cdot S(\omega,\omega')$, where $w_A(\omega), w_B(\omega)$ are the weight of ω in A and B. and $|A| = \sum_\omega w_A(\omega)$ and $|B| = \sum_\omega w_B(\omega)$.

B.1 Section 2 Proofs

Proof of Lemma 1 and Proposition 1 The agent's assessment of $E[\psi]$ is given by:

$$\begin{aligned} E_s[\psi(\omega)|C] &= \frac{\sum_\omega \pi(\omega)\cdot S(\omega,C)\cdot\psi(\omega)}{\sum_\omega \pi(\omega)\cdot S(\omega)} = \frac{E_\pi[S(\omega,C)\cdot\psi(\omega)]}{E_\pi[S(\omega)]} \\ &= \frac{Cov_\pi[S(\omega,C),\psi(\omega)] + E_\pi[S(\omega,C)]E_\pi[\psi(\omega)]}{E_\pi[S(\omega,C)]} \\ &= \frac{Cov_\pi[S(\omega,C),\psi(\omega)]}{E_\pi[S(\omega,C)]} + E[\psi(\omega)], \end{aligned} \tag{26}$$

where E_π refers to the expectation taken over ω with measure $\pi(\omega)$. Thus, we obtain:

$$E_s[\psi(\omega)|C] - E[\psi(\omega)] = \frac{Cov_\pi[S(\omega,C),\psi(\omega)]}{E_\pi[S(\omega,C)]}. \tag{27}$$

To conclude, recall $S(\omega,C) = \sum_{\omega_c\in C} w(\omega_c)\cdot S(\omega,\omega_c)$, which implies $E_\pi[S(\omega,C)] = \mathcal{W}(C) = \sum_{\omega_c\in C} w(\omega_c)\cdot\mathcal{W}(\omega_c)$, and by the linearity of the covariance function, one obtains:

$$E_s[\psi(\omega)|C] - E[\psi(\omega)] = \frac{\sum_{\omega_c\in C} w(\omega_c)\cdot Cov_\pi[S(\omega,\omega_c),\psi(\omega)]}{\sum_{\omega_c\in C} w(\omega_c)\cdot\mathcal{W}(\omega_c)} = \frac{\sum_{\omega_c\in C} w(\omega_c)\cdot\mathcal{W}(\omega_c)\cdot D^\psi(\omega_c)}{\sum_{\omega_c\in C} w(\omega_c)\cdot\mathcal{W}(\omega_c)}, \tag{28}$$

as desired. The specific form in Lemma 1 can be easily obtained from setting $\psi(\omega) = I(\omega\in H)$, and observing that:

$$\mathcal{D}^\psi(\omega_c) = \frac{\sum_{\omega\in H} S(\omega,\omega_c)\pi(\omega)}{\mathcal{W}(\omega_c)} - \pi(H) \tag{29}$$

B.2 Section 4 Proofs

Proof of Proposition 2 Optimization of the similarity-weighted intertemporal utility yields the following indifference condition between consumption at time t and consumption at time $t+1$:

$$S(\omega_t, C_t) \cdot u'(c_t) = \beta \cdot R \cdot \left[\sum_{\omega_{t+1}} \pi(\omega_{t+1}|\omega_t) \cdot S(\omega_{t+1}, C_t) \cdot u'(c_{t+1}, \omega_{t+1}) \right]. \tag{30}$$

Rearranging, one obtains:

$$S(\omega_t, C_t) = \beta \cdot R \cdot E_t\left[S(\omega_{t+1}, C_t) \cdot M(\omega_{t+1})\right] = \beta \cdot R \cdot E_t[S(\omega_{t+1}, C_t)] \cdot \left[\mathcal{D}^M(C_t) + E_t[M(\omega_{t+1})]\right]. \tag{31}$$

Dividing both sides by $E_t[S(\omega_{t+1}, C_t)]$ and rearranging, we obtain:

$$\begin{aligned} \beta \cdot R \cdot E_t[M(\omega_{t+1})] &= -\beta \cdot R \cdot \mathcal{D}^M(C_t) + S(\omega_t, C_t) \cdot \left(E_t[S(\omega_{t+1}, C_t)]\right)^{-1} \\ &= -\beta \cdot R \cdot \mathcal{D}^M(C_t) + \left(E_t[\tilde{S}(\omega_{t+1}, C_t)]\right)^{-1} \end{aligned} \tag{32}$$

Proof of Proposition 3 Suffices to compute for this specification the expected similarity of future states ω_{t+h} to the current state ω_t. In particular, the effective weight of time $t+h$ utility (net of the rational discount factor β^h) when the agent optimizes at time t in state ω_t is given by:

$$E_t[S(\omega_{t+h}, \omega_t)] = \int \exp\left(-\frac{\kappa}{2}(\omega_t - \omega_{t+h})^2\right) dF(\omega_{t+h}), \tag{33}$$

where the distribution of $dF(\omega_{t+h})$ given ω_t is given by $\mathcal{N}\left(\rho^h \omega_t, \frac{1-\rho^{2h}}{1-\rho^2}\sigma^2\right)$. To compute this quantity, the following lemma is useful:

Lemma 2. *Suppose the similarity function is given by $S(\omega, \omega') = \exp\left(-\frac{\kappa(\omega-\omega')^2}{2}\right)$, and states are normally distributed: $\phi_{\mathcal{N}}(\omega; \mu, \sigma^2)$. Then, $S(\omega, x^*) \cdot \phi_{\mathcal{N}}(\omega; \mu, \sigma^2))$ is exactly equal to $\xi((\mu - x^*)^2, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\frac{\mu + \kappa\sigma^2 x^*}{1+\kappa\sigma^2}, \frac{\sigma^2}{1+\kappa\sigma^2}\right)$, where $\xi((\mu - x^*)^2, \kappa, \sigma^2) \equiv \frac{1}{\sqrt{1+\kappa\sigma^2}} \cdot \exp\left(-\frac{\kappa(\mu-x^*)^2}{2(1+\kappa\sigma^2)}\right)$.*

Proof. The expression follows from the following standard algebraic manipulation:

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{\kappa(x-x^*)^2}{2}\right) &= \exp\left(-\frac{\kappa(\mu - x^*)^2}{2(1+\kappa\sigma^2)}\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(1+\kappa\sigma^2)}{2\sigma^2}\left(x - \frac{\mu + \kappa\sigma^2 x^*}{1+\kappa\sigma^2}\right)^2\right) \\ &= \xi((\mu - x^*)^2, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\frac{\mu + \kappa\sigma^2 x^*}{1+\kappa\sigma^2}, \frac{\sigma^2}{1+\kappa\sigma^2}\right). \end{aligned} \tag{34}$$

□

In particular, the lemma implies that the resonance of cue x^* is given by $\frac{1}{\sqrt{1+\kappa\sigma^2}} \cdot$

$\exp\left(-\frac{\kappa(\mu-x^*)^2}{2(1+\kappa\sigma^2)}\right) = \sqrt{\frac{\tau}{\tau+\kappa}} \cdot \exp\left(-\frac{\kappa(\mu-x^*)^2}{2(1+\kappa\sigma^2)}\right)$, where $\tau = (\sigma^2)^{-1}$. Applying this lemma directly yields:

$$E_t[S(\omega_{t+h}, \omega_t)] = \sqrt{\frac{\tau}{\tau+\kappa}} \cdot \exp\left(-\frac{1}{2}\left(\frac{\tau \cdot \kappa}{\tau+\kappa} \cdot (1-\rho^h)^2 \omega^2\right)\right), \tag{35}$$

where $\tau = \frac{1-\rho^2}{\sigma^2(1-\rho^{2(h+1)})}$, as desired.

Proof of Proposition 4 We break down the proof into two prerequisite lemmas. For the following exercise, denote $\gamma \equiv \frac{w_{Q^*}(\omega_A)+w_{Q^*}(\omega_B)}{w(\omega_0)\mathcal{W}(\omega_0)}$, and denote $\gamma_1 = \frac{w_{Q^*}(\omega_A)}{w(\omega_0)\mathcal{W}(\omega_0)}$ and $\gamma_2 = \frac{w_{Q^*}(\omega_B)}{w(\omega_0)\mathcal{W}(\omega_0)}$, where $\gamma = \gamma_1 + \gamma_2$

Proof. We require the following lemmas:

Lemma 3. *Denote $\pi_A \equiv \sum_{\omega \in H_A} \pi(\omega|C,Q)$ and $\pi_B \equiv \sum_{\omega \in H_B} \pi(\omega|C,Q)$. The set of beliefs that the persuader can induce is given by (π_A, π_B) such that:*

$$\pi_A + \pi_B = \frac{2\delta_0 + \gamma(1+\delta_1)}{1+2\delta_0+\gamma(1+\delta_0+\delta_1)}p, \ \min\{\pi_A, \pi_B\} \geq \frac{\delta_0 + \delta_1\gamma}{1+2\delta_0+\gamma(1+\delta_0+\delta_1)}p \tag{36}$$

Proof. The lemma follows from the fact that the beliefs induced by the cue is given by:

$$(\pi_1, \pi_2, \pi_0) \propto (\gamma_1 + \gamma_2\delta_1 + \delta_0, \gamma_2 + \gamma_1\delta_1 + \delta_0, 1 + \gamma\delta_0). \tag{37}$$

□

Lemma 4. *Suppose that the persuader is constrained to offer a portfolio with fixed weight $(x_1, x_2) = W \cdot (\eta_1, \eta_2)$, where $\eta_1 + \eta_2 = 1$ and $\eta_i \geq 0$. The persuader's message consists of a pure cue: $(\gamma_1, \gamma_2)^* = (\gamma, 0)$ if $\eta_1 > \eta_2$ and $(0, \gamma)$ otherwise.*

Proof. Assume without loss of generality that $\eta_1 \geq \eta_2$. First, we solve for the mean and variance of the returns given beliefs (π_0, π_1, π_2). We apply the law of iterated covariance:

$$Cov(X,Y) = E[Cov(X,Y|Z)] + Cov[E[X|Z], E[Y|Z]], \tag{38}$$

where X and Y are R_i, and Z is the realization of the stochastic state ω. One can work out that given the above set-up, the perceived covariance of returns for $[R_1, R_2]$ is given by:

$$E[R] = \alpha \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}, \ Var[R] = \begin{bmatrix} \sigma^2 + \alpha^2\pi_1(1-\pi_1) & -\alpha^2\pi_1\pi_2 \\ -\alpha^2\pi_1\pi_2 & \sigma^2 + \alpha^2\pi_2(1-\pi_2), \end{bmatrix} \tag{39}$$

Consider a portfolio with weight η_i in asset i with $\eta_1 + \eta_2 = 1$. The maximal fee for these relative weights is given by:

$$
\begin{aligned}
f^*(\eta_1,\eta_2,\pi_1,\pi_2) &= \max_W W(\eta_1\pi_1+\eta_2\pi_2)\alpha - \frac{A}{2}W^2\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}Var[R]\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \frac{1}{2A}\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}Var[R]\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}}\\
&= \frac{1}{2A}\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{(\sigma^2+\alpha^2\pi_1(1-\pi_1))\eta_1^2+(\sigma^2+\alpha^2\pi_2(1-\pi_2))\eta_2^2-2\alpha^2\pi_1\pi_2\eta_1\eta_2}\\
&= \frac{1}{2A}\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{\sigma^2(\eta_1^2+\eta_2^2)+(\pi_1\eta_1^2+\pi_2\eta_2^2)\alpha^2-(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}
\end{aligned}
\tag{40}
$$

Given that $f(x) = 1/(x^{-1}+1)$ is monotonically increasing in x, $f^*(\eta_1,\eta_2)$ is maximized if and only if:

$$
\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{\sigma^2(\eta_1^2+\eta_2^2)+(\pi_1\eta_1^2+\pi_2\eta_2^2)\alpha^2}
\tag{41}
$$

is maximized.

By Lemma 3, suffices to consider the case in which $\pi_1+\pi_2 = \bar{\pi}$, with π_1, π_2 ranging over $\bar{\pi}\cdot(1-\pi^*, \pi^*)$ where $\pi^* > \frac{1}{2}$. We proceed with the proof in two steps. First, we prove that the expression in Equation 41 is U-shaped: increasing for $\pi_1 > C$ and decreasing for $\pi_1 < C$. Second, we then compare expression for the lowest and highest possible values of π_1: $\pi_1 = \bar{\pi}\pi^*$ and $\bar{\pi}(1-\pi^*)$ to show that the former is higher. Put together, this implies that the expression is globally maximized at π_1 at its maximal value (which corresponds to $(\gamma_1,\gamma_2) = (\gamma, 0)$.).

Substituting in $\pi_2 = \bar{\pi} - \pi_1$, and taking the derivative with respect to π_1, we obtain that the expression is positive if and only if:

$$
\begin{aligned}
&2(\eta_1-\eta_2)((\eta_1-\eta_2)\pi_1+\eta_2\bar{\pi})\cdot(\kappa(\eta_1^2+\eta_2^2)+(\bar{\pi}\eta_2^2+(\eta_1^2-\eta_2^2)\pi_1))-((\eta_1-\eta_2)\pi_1+\eta_2\bar{\pi})^2\cdot(\eta_1^2-\eta_2^2)>0\\
&\iff 2\kappa(\eta_1-\eta_2)(\eta_1^2+\eta_2^2)+2(\eta_1-\eta_2)(\eta_1^2\pi_1+\eta_2^2(\bar{\pi}-\pi_1))-(\eta_1\pi_1+\eta_2(\bar{\pi}-\pi_1))(\eta_1^2-\eta_2^2)>0
\end{aligned}
\tag{42}
$$

The expression is linear in π_1 with the slope of π_1 given by $(\eta_1-\eta_2)(\eta_1^2-\eta_2^2) > 0$: thus, the expression is U-shaped (which includes monotonically increasing or decreasing). Thus, to find the global maximum, suffices to compare the original function evaluated at the end points: $\pi_1 = \pi_1^* > \frac{1}{2}$ and $\pi_1 = 1-\pi_1^* = \pi_2^*$. Thus, suffices to show:

$$\frac{(\eta_1\pi_1^*+\eta_2\pi_2^*)^2}{\kappa(\eta_1^2+\eta_2^2)+(\eta_1^2\pi_1^*+\eta_2^2\pi_2^*)} > \frac{(\eta_1\pi_2^*+\eta_2\pi_1^*)^2}{\kappa(\eta_1^2+\eta_2^2)+(\eta_1^2\pi_2^*+\eta_2^2\pi_1^*)} \tag{43}$$

Expanding and using the fact that $\eta_1\pi_1^*+\eta_2\pi_2^* \geq \eta_1\pi_2^*+\eta_2\pi_1^*$, suffices to show:

$$(\eta_1^2\pi_2^*+\eta_2^2\pi_1^*)(\eta_1\pi_1^*+\eta_2\pi_2^*)^2-(\eta_1^2\pi_1^*+\eta_2^2\pi_2^*)(\eta_1\pi_2^*+\eta_2\pi_1^*)^2 = \pi_1^*\pi_2^*(\pi_1^*-\pi_2^*)(\eta_1^2-\eta_2^2)(\eta_1-\eta_2)^2 > 0, \tag{44}$$

as desired. □

To conclude the proof, by Lemma 4, suffices to assume without loss of generality that $(\gamma_1,\gamma_2)=(\gamma,0)$. Let (π_1,π_2) be the induced beliefs given that cue. One can easily see that the optimal portfolio given (π_1,π_2) is given by:

$$x^* = Var[R]^{-1}E[R] = \frac{1}{(\sigma^2+\alpha^2\pi_1(1-\pi_1))(\sigma^2+\alpha^2\pi_2(1-\pi_2))-\alpha^4\pi_1^2\pi_2^2}\begin{bmatrix}\sigma^2\pi_1+\alpha^2\pi_1\pi_2\\ \sigma^2\pi_2+\alpha^2\pi_1\pi_2\end{bmatrix}, \tag{45}$$

with the tilt of the portfolio given by

$$T(x^*) = \frac{\sigma^2\pi_1+\alpha^2\pi_1\pi_2}{\sigma^2\pi_2+\alpha^2\pi_1\pi_2}. \tag{46}$$

Denote for simplicity $\kappa \equiv \frac{\alpha^2}{\sigma^2}$. One can derive that $\frac{\partial T}{\partial \delta_1} < 0$ if and only if:

$$\pi_2(1+\kappa\pi_2)\frac{\partial \pi_1}{\partial \delta_1} < \pi_1(1+\kappa\pi_1)\frac{\partial \pi_2}{\partial \delta_1} \tag{47}$$

Recall that:

$$\pi_1 = \frac{\gamma+\delta_0}{(1+2\delta_0)+\gamma(1+\delta_0+\delta_1)}, \pi_2 = \frac{\gamma\delta_1+\delta_0}{(1+2\delta_0)+\gamma(1+\delta_0+\delta_1)}, \tag{48}$$

and in particular $\frac{\partial \pi_1}{\partial \delta_1} < 0$ and $\frac{\partial \pi_2}{\partial \delta_1} > 0$. This automatically implies the above inequality.

□

Proof of Proposition 5 Recall that we assume the following similarity structure: $S(\omega,\omega') = \exp\left(-\frac{\kappa}{2}(\omega-\omega')^2\right)$. Note that given any normal density $\phi_{\mathcal{N}}(\mu_0,\tau_0^{-1})$, the density distorted by the cue Q is proportional to (excluding all terms unrelated to μ_0 and

Q):

$$\exp\left[-\frac{\tau_0}{2}(\omega-\mu_0)^2-\frac{\kappa}{2}(\omega-Q)^2\right]=\exp\left[-\frac{\tau_0+\kappa}{2}\left(\omega-\frac{\tau_0\mu_0+\kappa Q}{\tau_0+\kappa}\right)^2-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right]. \quad (49)$$

This implies that for a cue (C,Q), the subjective distribution is a mixture of two normal distributions $\phi_{\mathcal{N}}\left(\frac{\tau_0\mu_0+\kappa Q}{\tau_0+\kappa},(\tau_0+\kappa)^{-1}\right)$, $\phi_{\mathcal{N}}\left(\frac{\tau_0\mu_0+\kappa\mathcal{D}(C)}{\tau_0+\kappa},(\tau_0+\kappa)^{-1}\right)$ each with weight proportional to $\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right)$ and $\mathcal{W}(C)\cdot\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-\mathcal{D}(C))^2}{\tau_0+\kappa}\right)$. As assumed in the proposition, we focus on the case where $\mathcal{D}(C)=\mu_{post}$: the context absent the persuader does not on its own introduce belief distortion. Then, the subjective beliefs simplify to:

$$E_s[\omega]=\mu_0+\frac{\gamma\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right)}{1+\gamma\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right)}\cdot\frac{\kappa}{\tau_0+\kappa}(Q-\mu_0), \quad (50)$$

where $\gamma=(\mathcal{W}(C))^{-1}$. Setting $\eta_0\equiv\frac{\kappa\tau_0}{\tau_0+\kappa}$ and $x\equiv Q-\mu_0$, it suffices to maximize (in x) the expression:

$$-\frac{1}{2}\eta_0x^2-\log\left(1+\gamma\exp\left(-\frac{1}{2}\eta_0x^2\right)\right)+\log(x), \quad (51)$$

which generates the FOC:

$$0=-\eta_0x+\frac{1}{x}+\frac{\gamma\exp\left(-\frac{1}{2}\eta_0x^2\right)\cdot\eta_0x}{1+\gamma\exp\left(-\frac{1}{2}\eta_0x^2\right)}\implies 1=\frac{1}{1+\gamma\exp\left(-\frac{1}{2}\eta_0x^2\right)}\eta_0x^2, \quad (52)$$

which is satisfied when $\eta_0x^2=\nu_0$. Thus, $x=\sqrt{\nu_0/\eta_0}=\sqrt{\frac{\tau_{post}+\kappa}{\tau_{post}\kappa}\nu_0}k$, as desired. The expression for $E[\omega|Q^*,I]$ is immediate from substituting this expression into Equation 50. In particular, given that $\sqrt{\frac{\kappa}{\tau_{post}\cdot(\tau_{post}+\kappa)}}$ is decreasing in τ_{post}, we immediately obtain that optimal beliefs are decreasing in information precision.

C Informative persuasion benchmark

Consider the case where the persuader can provide true information at a cost, with the costs proportional to the mutual information between the prior and the posterior (Sims 2011). Formally, the informative persuader maximizes over local perturbations $(\pi(\omega_A),\pi(\omega_B)) = (\pi_0(\omega_A)+d\pi(\omega_A),\pi_0(\omega_B))+d\pi(\omega_B))$ as follows:

$$\max \phi^*(\pi(\omega_A),\pi(\omega_B)) - \xi\left[\log\left(\frac{\pi(\omega_A)}{\pi_0(\omega_A)}\right)\pi(\omega_A) + \left(\frac{\pi(\omega_B)}{\pi_0(\omega_B)}\right)\pi(\omega_B) + \left(\frac{\pi(\omega_0)}{\pi_0(\omega_0)}\right)\pi(\omega_0)\right], \quad (53)$$

where $\phi^*(\pi(\omega_A),\pi(\omega_B))$ are optimal fees given investor beliefs. In that case, the persuader instead chooses the following strategy:

Proposition 6. *The informative persuader locally boosts π_A and π_B equally and offers a perfectly diversified portfolio: $T(x^*) = 1$.*

Proof. Recall that the cost of moving someone's beliefs (through information) from $(\pi_1^d,\pi_2^d,\pi_0^d)$ to (π_1,π_2,π_0) is given by $\xi\cdot D_{KL}(\pi\|\pi_d)$. Given our assumption that ξ is sufficiently high, suffices to consider marginal perturbations of beliefs near $(\pi_1^d,\pi_2^d,\pi_0^d)$.

We have from the above derivation that the optimal fees given beliefs π_1,π_2 is given by:

$$F(\pi_1,\pi_2) \equiv \frac{\alpha^2(\pi_1^2+\pi_2^2)\sigma^2+\alpha^2(\pi_1+\pi_2)\pi_1\pi_2}{\sigma^4+\alpha^2(\pi_1+\pi_2)\sigma^2+\pi_1\pi_2\cdot\alpha^4-\alpha^2\left[(\pi_1^2+\pi_2^2)\sigma^2+\alpha^2(\pi_1+\pi_2)\pi_1\pi_2\right]}. \quad (54)$$

The informative persuader maximizes over local perturbations $\pi_1 = \pi_1^d + d\pi_1$ and $\pi_2 = \pi_2^d + d\pi_2$ the following expression:

$$\max F(\pi_1^d+d\pi_1,\pi_2^d+d\pi_2) - \xi\left[\log\left(\frac{\pi_1}{\pi_1^d}\right)\pi_1 + \log\left(\frac{\pi_2}{\pi_2^d}\right)\pi_2 + \log\left(\frac{\pi_0}{\pi_0^d}\right)\pi_0\right]. \quad (55)$$

The persuader objective admits the following Taylor approximation:

$$\nabla_1 F d\pi_1 + \nabla_2 F d\pi_2 - \frac{\xi}{2}\left(\frac{d\pi_1^2}{\pi_1^d}+\frac{d\pi_2^2}{\pi_2^d}+\frac{(d\pi_1+d\pi_2)^2}{\pi_0^d}\right), \quad (56)$$

where $\nabla_1 F = \nabla_2 F \equiv \nabla > 0$. Putting this together, for sufficiently large ξ, the following FOC should hold:

$$0 = \nabla - \xi\cdot\frac{1}{\pi_1^d}d\pi_1 - \xi\frac{d\pi_1+d\pi_2}{\pi_0^d} \implies d\pi_1^* = \frac{\nabla}{\xi}\cdot\left(\frac{1}{\pi_1^d}+\frac{1}{\pi_0^d}\right) = d\pi_2^*, \quad (57)$$

as desired. □