Decomposing Belief Updating: Evidence Evaluation and Posterior Formation*

Tashfeen Saeed†

October 26, 2026

Abstract

While a large empirical literature documents systematic biases in belief updating, previous work elicits posterior beliefs directly, leaving unresolved at which stage these biases arise. Using an experimental framework, I separate and examine the two stages of belief updating: evaluating evidence and forming posteriors with that evidence. I show that systematic distortions in the form of overestimation begin during evidence evaluation before it is combined with priors. These distortions account for much of posterior underinference: reported posteriors closely track those obtained by applying Bayes' rule to participants' own likelihood reports, while independently elicited likelihoods reproduce much of the posterior bias in a posterior-only baseline. Intervening at the evidence evaluation stage – either through elicitation or information provision – additionally improves posterior beliefs. In a second experiment, I show that absolute likelihood levels systematically shape likelihood judgments and posterior beliefs even when the objective likelihood ratio remains fixed, violating a key ratio-invariance restriction shared by Bayesian updating and a wide class of behavioral models. These findings establish the importance of evidence evaluation in addressing systematic belief biases and in informing richer models of belief formation.

*I am grateful to Ted O'Donoghue, Lawrence Jin, Collin Raymond, and Sam Hirshman for guidance and support. I thank Daniel Benjamin, Devin Pope, Pietro Ortoleva, Nicolas Bottan, David Huffman, Johannes Haushofer, Michèle Belot, and the participants at Cornell BERG for discussions or comments that improved this paper. Both experiments in this paper were preregistered with the AEA RCT Registry (AEARCTR-0017734, AEARCTR-0019513) and received exemption determinations from Cornell University's Office of Research Integrity and Assurance (IRB0150154, IRB0151094).

†Department of Economics, Cornell University, 455 Uris Hall, Ithaca, NY 14853. Email: tfs44@cornell.edu.

1 Introduction

Economic decision-making relies on beliefs about uncertain states. Individuals update these beliefs based on new signals—doctors assess patient health from medical tests, employers learn about applicants from résumé credentials, and investors infer firm profitability from earnings announcements. While a large empirical literature documents systematic biases in these beliefs across laboratory and field settings (e.g., Benjamin, 2019; Bohren et al., 2019; Mullainathan and Obermeyer, 2022), most previous work observes posterior beliefs directly, leaving unresolved where in the updating process these biases arise.

Conceptually, forming a posterior belief under Bayesian updating involves two distinct stages: (i) *evidence evaluation*, assessing likelihoods for the observed signals under each possible state, and (ii) *posterior formation*, using Bayes' rule to map these likelihoods into posterior beliefs. Consider a doctor diagnosing a patient's health after observing a series of positive test results. The doctor must first evaluate how likely these results are if the patient is healthy versus if the patient is ill, then combine that evaluation with prior patient health beliefs to form a posterior.[1] A biased diagnosis can therefore stem from errors in the doctor's evaluation of the evidence, in combining that evaluation with prior beliefs, or both. Because the final posterior belief reflects the output of both stages, observing it alone cannot reveal which source of error needs to be addressed to correct beliefs.

In this paper, I make the evidence evaluation stage observable and examine its contribution to belief biases. By experimentally separating the two stages, I show that participants systematically overestimate likelihoods before posteriors are formed. These distortions account for much of posterior under-inference: reported posteriors closely track those obtained by using Bayes' rule on reported likelihoods, both within participants and across independent treatments. Reporting or providing likelihoods further improves posterior beliefs and closes approximately 18–20 percent of the gap to the Bayesian benchmark. These documented likelihood distortions additionally suggest a novel channel for how absolute likelihood levels shape belief biases, which I examine in a second experiment. Holding relative support fixed, I show that lower likelihood levels generate greater likelihood overestimation and posterior under-inference, revealing a systematic dependence that Bayesian updating and a large group of behavioral models do not capture. These findings establish the importance of the evidence evaluation stage in addressing systematic belief biases and in informing richer models of belief formation.

To separate the two stages in a controlled environment, I build on the widely used bookbag-

[1]These two stages are directly relevant in diagnostic medicine, where these inputs correspond to the false-positive and true-positive probabilities, respectively, and the patient's pre-test probability of illness. Disease prevalence among comparable patients informs the latter.

and-poker-chip paradigm, which provides exact benchmarks for each stage of Bayesian updating. In each round, there are two bags containing green and orange balls in known but different proportions. One bag is selected at random, and participants observe a sample of balls drawn from that bag. Participants then make two distinct probability judgments corresponding to each stage of updating: the *likelihood* of drawing the observed sample from each bag and the *posterior belief* that each bag generated the sample. Across four treatment groups, I vary which of these judgments are reported or provided: participants either (i) report posterior beliefs only, as in most prior studies, (ii) report both likelihoods and posterior beliefs, (iii) are provided with likelihoods and report posterior beliefs, or (iv) report likelihoods only. This design isolates the two stages of updating and allows me to examine how distortions propagate across them.

The experimental results first confirm the patterns documented in past work. In the posterior-only baseline treatment, participants under-infer on average, revising their posterior beliefs to about 69 percent of the Bayesian benchmark. However, a stark contrast emerges at the evidence evaluation stage. In treatments where participants report likelihood judgments—either alone or jointly with posterior beliefs—reports closely track objective benchmarks for single signals but become systematically distorted upward for multiple signals, with large estimated positive intercepts ranging from 19–24 percentage points. These substantial upward distortions arise even in the likelihood-only treatment, showing that systematic errors arise at the evidence evaluation stage without any requirement to form a posterior belief.

Next, these distorted likelihood judgments account for much of the resulting posterior under-inference, both within-treatment and across treatments. Among participants who report both likelihoods and posteriors, reported beliefs very closely track those obtained by applying Bayes' rule to participants' own likelihood judgments, with little additional systematic attenuation. Importantly, likelihood judgments from the likelihood-only treatment reproduce much of the posterior under-inference in the posterior-only baseline group. Participants asked only to evaluate evidence therefore report likelihoods that, when combined through Bayes' rule, recover much of the belief bias among participants asked only to report posteriors, showing that this relationship cannot be explained by jointly reporting likelihoods and posteriors. These findings locate an important source of posterior bias in evidence evaluation, consistent with work in which individuals combine distorted subjective evidence approximately according to Bayes' rule (e.g., Rabin, 2002, 2013).

The randomized treatment comparisons further show that intervening at this intermediate stage improves belief updating. Relative to the baseline adjustment of 69 percent, adjustments increase to about 75 percent (a 9% increase) both when participants additionally report likelihood judgments and when objective likelihoods are provided. Both adjustments are

significantly larger than in the posterior-only treatment but not statistically different from each other, indicating that explicitly engaging the likelihood stage through reporting can improve belief updating even without providing correct likelihoods. To interpret this result through the earlier example, a doctor who first explicitly evaluates or is given the likelihood of the test results would form a more accurate diagnosis by closing roughly 18–20 percent of the gap relative to relying on intuition alone. These results are robust even after excluding participants who do not meet a pre-registered set of sample filters, including comprehension check and completion time requirements, which suggests that incorrect comprehension of the task or inattention cannot account for the findings.

The results so far identify evidence evaluation as an important source of posterior bias, but a closer look across information environments reveals another puzzle: substantial likelihood overestimation accompanies both relatively accurate posteriors and pronounced under-inference across different environments. To understand this variation, Bayesian updating provides a framework for how likelihood judgments translate into beliefs: with fixed priors, the posterior depends solely on the *relative* support for the two states, summarized by the ratio of the two likelihoods. A broad class of behavioral models preserves this restriction while allowing a stable compression or amplification of the objective likelihood ratio, commonly estimated through a parameter (e.g., Edwards, 1968; Grether, 1980; Aina et al., 2023; Gonçalves et al., 2026).

The upward shifts documented in reported likelihood judgments, however, suggest a novel channel for posterior bias: *absolute likelihood levels*. A common additive upward distortion in both likelihoods would make the two more similar in proportional terms, moving their ratio toward one and weakening relative evidence. This effect would be stronger when both objective likelihoods are small. If participants combine these judgments approximately according to Bayes' rule, lower likelihood levels would therefore generate greater posterior under-inference even though objective relative evidence is unchanged, departing from predictions of Bayesian updating and ratio compression models. However, this channel is difficult to isolate in the main experiment because the symmetric bag structure links variation in likelihood levels to changes in relative evidentiary strength, sample size, and composition.

To isolate this channel, in a second experiment, I construct matched environments that vary absolute likelihood levels while holding the objective likelihood ratio, sample size, and sample composition fixed. Asymmetric bag compositions make the same observed sample proportionally more or less likely under both states without changing its relative evidentiary strength or Bayesian posterior. Because two environments with different absolute likelihoods still have the same likelihood ratio, the evidence provides the same relative support for one state over the other. Bayesian updating and ratio-based behavioral models therefore treat

these environments as informationally equivalent and predict the same posterior belief within a fixed ratio.

Contrary to this prediction, the results show that the underlying likelihood level systematically affects both reported likelihood judgments and posterior beliefs. Lower absolute likelihood levels generate greater likelihood overestimation and posterior underinference within identical ratios, with the bias declining systematically as the absolute likelihood level grows larger. The subjective likelihood ratios reveal an important distinction: median ratios closely match the objective benchmark for single draws, but are compressed toward one for two-draw samples, with less compression at higher likelihood levels. In some conditions, increasing likelihood levels reverses posterior beliefs from underinference to overinference despite no change in the relative evidence. These results are difficult to reconcile with a stable transformation of likelihood ratios, which is widely used in this literature, and show that understanding belief formation requires a deeper understanding of how individuals evaluate the evidence itself—a stage that a broad class of behavioral models largely treat as given.

This paper is related to three strands of literature. First, it is directly connected to the large body of work in psychology and economics on biases in belief formation (e.g., Tversky and Kahneman, 1971; Camerer, 1987; Grether, 1992; Griffin and Tversky, 1992; Holt and Smith, 2009). As summarized by Benjamin (2019), a central finding in this literature is under-inference: individuals update their beliefs less than Bayesian benchmarks predict.[2] More recent work investigates the cognitive foundations of these deviations (e.g., Enke and Graeber, 2023; Ba et al., 2024; Augenblick et al., 2025), documenting over-inference in weak, noisy, or complex environments. However, because these studies typically elicit posterior beliefs directly, they cannot distinguish distorted evidence evaluation from errors in posterior formation. This paper is the first to separately measure and connect these stages within a common experimental framework. By doing so, it reveals how these documented deviations arise at an earlier and previously unobserved stage with distorted likelihood inputs, which account for a significant portion of posterior under-inference. Additionally, it shows that eliciting or providing likelihoods improves posterior updating, highlighting the effectiveness of interventions at this intermediate stage to reduce the biases documented in this literature.

Second, the paper contributes to the literature on how individuals evaluate and represent evidence. A large body of work explains belief biases through distortions in the underlying evidence, including sampling distribution beliefs, representativeness, and selective attention (e.g., Rabin, 2002; Rabin and Vayanos, 2010; Gennaioli and Shleifer, 2010; Benjamin et al.,

[2]In contrast to insufficient responsiveness to signals, a related line of work documents *base-rate neglect*, or incorrect use of prior information (e.g., Kahneman and Tversky, 1973; Bar-Hillel, 1980; Esponda et al., 2024). Because such distortions arise in settings with unequal priors, I focus on updating problems with equal priors (50–50) to isolate distortions in inference.

2016; Bordalo et al., 2026). In parallel, related experiments document distortions in judgments about signal frequencies, compound events, and sampling distributions (e.g., Kahneman and Tversky, 1972; Bar-Hillel, 1973; Tversky and Kahneman, 1983; Khaw et al., 2017; Benjamin et al., 2017; Prat-Carrabin and Woodford, 2024). However, there is little empirical evidence on how these evidence evaluation distortions systematically propagate into posterior beliefs, which are the inputs that govern decisions in most economics models. This paper reveals systematic distortions at this evaluation stage and shows how they translate to biases in posterior beliefs. It further identifies absolute likelihood levels as a systematic determinant of subjective evidence: even when the objective likelihood ratio is unchanged, likelihood judgments and posterior beliefs vary with the underlying likelihood levels. These findings identify a novel and overlooked feature of evidence that shapes its subjective representation and subsequent effect on posterior beliefs.

Third, this paper has implications for the literature on forecasting and expectations formation in macroeconomics and finance, which are studied both in the field (e.g., Coibion and Gorodnichenko, 2015; Giglio and Kelly, 2018; Bordalo et al., 2020, 2022) and in experiments (e.g., Afrouzi et al., 2023; Fan et al., 2024; Bastianello and Imas, 2026). In these settings, participants observe signals generated by an autocorrelated process and report forecasts or updated expectations about future realizations after each signal. This literature has documented systematic departures from rational expectations, including overreaction to recent signals, particularly at long horizons and for transitory processes. Although forecasting and belief updating have typically been studied separately, they are linked by the law of iterated expectations: a forecast can be expressed as an average of state-contingent expected outcomes, weighted by posterior beliefs about the underlying state. While this paper studies those posterior beliefs rather than forecasts directly, the findings suggest that forecast errors may partly originate in how individuals evaluate signals before forming posteriors, in addition to how they translate those posteriors into predictions. Separately measuring these steps could help distinguish sources of forecast bias that revisions alone leave unresolved.

The rest of the paper proceeds as follows. Section 2 presents a conceptual framework distinguishing evidence evaluation from posterior formation. Section 3 describes the design, treatments, and procedures of the main experiment. Section 4 reports the results, examining each stage separately and then connecting both. Section 5 presents the theoretical predictions, design, and results of the second experiment. Section 6 concludes.

2 Conceptual Framework

This section formalizes the distinction between evidence evaluation and posterior formation. I characterize the Bayesian benchmark for each stage, introduce subjective likelihoods, and define the benchmark for posterior formation conditional on an individual's subjective evaluation of the evidence. These distinctions provide the basis for the experimental design.

2.1 Bayesian Benchmark

Suppose there are two possible states of the world, A and B, which are equally likely a priori, $P(A) = P(B) = 0.5$. An individual observes a realized sample of signals S. To update prior beliefs to posterior beliefs about states, the individual must first evaluate how likely this sample of signals is under each possible state.

Definition 1 (Likelihood)**.** *The likelihoods of observing the realized signals S under each possible state are*

$$L_A \equiv P(S \mid A) \qquad \textit{and} \qquad L_B \equiv P(S \mid B). \tag{1}$$

The likelihood pair (L_A, L_B) summarizes how diagnostic the observed signals are for the two states. If $L_A > L_B$, the observed signals provide evidence in favor of state A; if $L_B > L_A$, they favor state B. To summarize the relative strength of this evidence, I next define the likelihood ratio.

Definition 2 (Likelihood Ratio)**.** *The likelihood ratio of the realized signals S is*

$$R \equiv \frac{L_A}{L_B}. \tag{2}$$

The likelihood ratio compares how much more likely the observed signals are under one state than the other. A likelihood ratio greater than one favors state A, whereas a likelihood ratio below one favors state B. Evidence becomes more diagnostic as $|\log R|$ increases: higher values of R provide stronger evidence for A, while lower values provide stronger evidence for B. Because R depends only on the relative magnitudes of L_A and L_B, likelihood pairs with different absolute levels can convey the same relative evidence about the state.

In the second stage, the individual combines the likelihoods with prior probabilities to form a posterior belief about the state. Under Bayes' rule and equal priors, this posterior depends on the likelihood pair only through its ratio.

Definition 3 (Posterior Belief)**.** *The posterior belief about state A after observing the realized signals S is*

$$P(A \mid S) = \frac{L_A}{L_A + L_B} = \frac{R}{1 + R}. \tag{3}$$

The posterior summarizes the individual's beliefs about the two states after incorporating the observed evidence and guides decision-making in economic models. Because the states are mutually exclusive and exhaustive, $P(B \mid S) = 1 - P(A \mid S)$. Posterior odds therefore satisfy

$$\frac{P(A \mid S)}{P(B \mid S)} = R,$$

so that under equal priors, the likelihood ratio in (2) translates directly into the relative posterior beliefs assigned to the two states.

A key implication is that the Bayesian posterior is invariant to the absolute levels of the likelihoods: scaling both likelihoods by a positive constant leaves the posterior unchanged. To illustrate this, $(L_A, L_B) = (0.80, 0.05)$ and $(L_A, L_B) = (0.08, 0.005)$ both imply $R = 16$ and hence the same Bayesian posterior. Although the sample is one-tenth as likely under either state in the second environment, it provides the same *relative* evidence about which state is true and leads to equivalent posteriors under Bayesian updating. Whether individuals also treat such environments as equivalent is a central question of this paper.

The Bayesian benchmark can therefore be decomposed as a two-stage mapping from the observed sample to likelihoods and from likelihoods to posterior beliefs:

$$S \xrightarrow{\text{evidence evaluation}} (L_A, L_B) \xrightarrow[\text{with prior beliefs}]{\text{posterior formation}} P(A \mid S). \tag{4}$$

Bayesian updating requires accurate evidence evaluation followed by posterior formation according to Bayes' rule. However, existing studies commonly observe only the final reported posterior, which jointly reflects both stages. A deviation from the Bayesian posterior therefore does not reveal whether the individual distorted the likelihoods when evaluating the evidence or combined them incorrectly when forming the posterior. The following subsection makes this distinction explicit by introducing subjective likelihoods.

2.2 Subjective Likelihoods and Posterior Formation

To distinguish deviations arising at each stage, I allow the individual's evaluation of the evidence to differ from the objective likelihoods.

Definition 4 (Subjective Likelihood)**.** *The subjective likelihoods of observing sample S under states A and B are*

$$\widetilde{L}_A \equiv \widetilde{P}(S \mid A) \qquad \textit{and} \qquad \widetilde{L}_B \equiv \widetilde{P}(S \mid B).$$

The corresponding subjective likelihood ratio is

$$\widetilde{R} \equiv \frac{\widetilde{L}_A}{\widetilde{L}_B}.$$

Subjective likelihoods capture the individual's evaluation of the observed evidence. Evidence evaluation is accurate when the subjective likelihoods coincide with their objective counterparts, $(\widetilde{L}_A, \widetilde{L}_B) = (L_A, L_B)$, and distorted otherwise.

After evaluating the evidence, the individual combines the subjective likelihoods with prior beliefs to form a subjective posterior. Under equal priors, the benchmark for this second stage can be defined conditional on the individual's subjective likelihoods.

Definition 5 (Conditional Bayesian Posterior Formation)**.** *Posterior formation is conditionally Bayesian if the individual applies Bayes' rule to the subjective likelihoods:*

$$\widetilde{P}(A \mid S) = \frac{\widetilde{L}_A}{\widetilde{L}_A + \widetilde{L}_B} = \frac{\widetilde{R}}{1 + \widetilde{R}}. \tag{5}$$

Conditional Bayesian posterior formation does not require evidence evaluation to be accurate. If subjective likelihoods differ from their objective counterparts, the individual may apply Bayes' rule correctly in the second stage and satisfy Equation (5), even though the resulting posterior may differ from the Bayesian benchmark in Equation (3).

The subjective updating process can now be summarized by the following two-stage mapping:

$$S \xrightarrow{\text{evidence evaluation}} (\widetilde{L}_A, \widetilde{L}_B) \xrightarrow[\text{with prior beliefs}]{\text{posterior formation}} \widetilde{P}(A \mid S). \tag{6}$$

Comparing this mapping with the Bayesian benchmark in (4) separates distortions into two potential sources. Evidence evaluation is distorted when subjective likelihoods depart from their objective counterparts in Definition 1; posterior formation is distorted when the resulting posterior differs from the conditional Bayesian posterior implied by those subjective likelihoods in Definition 5. Importantly, errors at each stage are distinct: inaccurate likelihoods may still be integrated correctly according to Bayes' rule. Therefore, observing both stages is important to understand how they contribute to errors.

This two-stage decomposition motivates the design of the main experiment to isolate distortions at each stage, as described in the next section.

3 Experiment 1: Decomposing Belief Updating

This section presents the design of the main experiment, which operationalizes the two-stage decomposition in Section 2.2. Across four randomized treatments, I vary whether likelihood judgments remain implicit, are elicited before posterior beliefs, or are provided to participants. This design makes the intermediate evidence-evaluation stage observable, traces its contribution to posterior errors, and identifies the effect of intervening at this stage on posterior beliefs.

3.1 Experimental Design

The experimental task builds on the standard balls-and-urns task widely used in the belief updating literature (Edwards, 1968). This task provides analytical benchmarks for both likelihood judgments and posterior beliefs which can be compared with reports at each stage to assess deviations. Additionally, it permits systematic variation of each stage and provides precise control over participants' information sets, which are often unobserved in field settings.

In each task, participants are shown two bags, A and B, each containing 100 green and orange balls. Let δ and $1-\delta$ denote the proportion of green balls in Bags A and B. Participants are told that each bag is equally likely to be selected. M balls are then drawn from the selected bag with replacement, of which G are green, and displayed in the order drawn.[3] Figure 1 illustrates an example task screen where $\delta = 0.75$, $M = 3$, and $G = 1$. In this example, Bag A contains 75 green balls and Bag B contains 25 green balls; participants observe three draws, of which one is green.

After observing the sample, participants report one or both of two probability judgments, depending on treatment assignment. These judgments correspond to the two stages of belief updating developed in Section 2. To reinforce the distinction, the instructions clarify the direction of conditioning:

1. **Likelihood judgments ("Sample Chance").** Participants report the probability of drawing the given sample separately from each bag. The bags are evaluated separately, so the two reports need not sum to 100.

2. **Posterior beliefs ("Which Bag").** Participants report the probability that Bag A generated the observed sample, where the probability assigned to Bag B is automatically calculated as 100 minus this report. The two reports are required to sum to 100.

[3] In treatments eliciting likelihoods only, participants are shown the bags and sample without mention of selecting one bag. The underlying randomization, information structure, and visual display are identical across treatments.

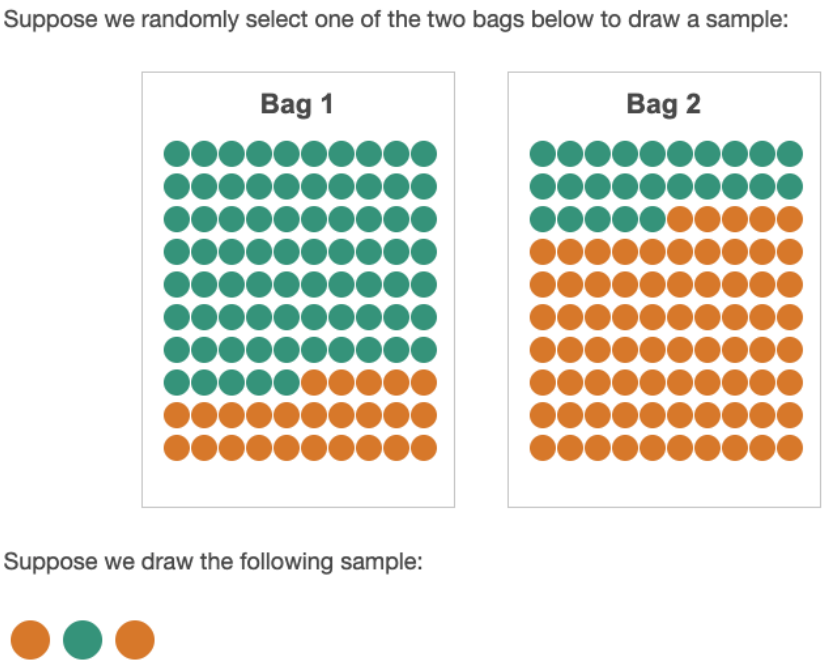


Figure 1. Experimental Interface, Main Experiment

Notes: Each bag contains 100 balls displayed as a 10×10 grid of colored circles. The sample consists of 1, 2, or 3 draws with replacement from the selected bag and is shown visually as an ordered sequence of colored balls. The experiment labels the bags as "Bag 1" and "Bag 2," corresponding to Bags A and B in the text.

All questions are framed hypothetically (e.g., "suppose a bag is selected" and "suppose the following sample is drawn"), allowing the objective likelihoods and Bayesian posterior benchmarks to vary exogenously with bag composition, sample size, and sample composition without deception.

3.2 Treatment Groups

To separately examine the two stages of Bayesian updating, participants are randomly assigned to one of four treatment groups that vary whether likelihoods (L) and posteriors (P) are reported or provided:

1. **Posteriors only (P):** Participants report posterior beliefs only; likelihoods are neither reported nor provided. This condition mirrors the standard belief-updating paradigm in the literature.

2. **Likelihoods and Posteriors (L/P):** Participants report likelihoods for each bag and posterior beliefs.

3. **Likelihoods only (L):** Participants report likelihoods for each bag only; posterior beliefs are not reported.

4. **Provided Likelihoods and Posteriors (L*/P):** Participants are provided with likelihoods for each bag and report posterior beliefs.

These four treatments allow complementary comparisons across the two stages. First, P replicates the standard posterior-elicitation design used in the existing literature to examine baseline posterior beliefs. Second, L/P makes intermediate likelihood judgments observable, allowing their relationship with reported posteriors to be examined both within treatment and across treatment with P. Third, L elicits likelihoods alone, making the first stage observable and examining whether it differs when reporting posteriors by comparing with L/P. Finally, L*/P provides objective likelihoods, allowing posterior formation to be examined when participants need not calculate these inputs themselves.

Figure 2 illustrates the experimental design and the treatment groups, while Appendix Figures E.1 to E.4 show the experimental task screen for each treatment condition.

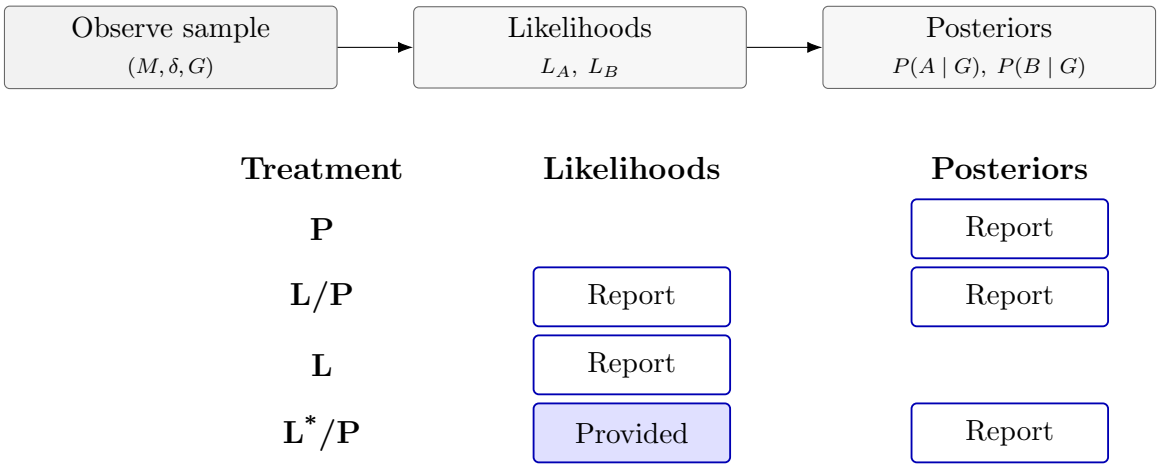


Figure 2. Experimental Design and Treatment Structure

Notes: Each round begins with an observed sample summarized by (M, δ, G), where M is the number of draws (signals) and G is the number of green draws. Likelihoods are denoted $L_A = P(G \mid A)$ and $L_B = P(G \mid B)$, the probabilities of drawing the observed sample from Bags A and B, respectively. Treatment labels indicate which objects are reported: P reports posteriors only; L/P reports likelihoods and posteriors; L reports likelihoods only; L*/P is provided likelihoods and reports posteriors.

3.3 Experimental Environments

For a particular ordered sequence of M draws containing G green balls, where δ and $1-\delta$ denote the proportion of green balls in Bags A and B, the objective likelihoods in Definition 1

are

$$L_A = \delta^G(1-\delta)^{M-G}, \qquad L_B = (1-\delta)^G\delta^{M-G}. \tag{7}$$

These likelihoods refer to the exact displayed sequence: for example, green followed by orange has likelihood $\delta(1-\delta)$ under Bag A. Because draws are independent and made with replacement, sequences with the same composition have equal likelihoods. Consequently, M, G, and δ are sufficient to calculate each sequence's likelihood.

Under equal prior probabilities, these likelihoods then determine the Bayesian posterior according to Equation (3). These expressions therefore provide exact analytical benchmarks for each stage of updating, allowing deviations in likelihood formation and posterior integration to be evaluated separately.

Each participant completes 15 tasks. In each task, a combination of bag composition, sample size, and sample composition is randomly selected according to

$$\delta \in \{0.65, 0.70, 0.75\}, \quad M \in \{1, 2, 3\}, \quad G \in \{0, \ldots, M\},$$

with one combination realized in each task (e.g., $\delta = 0.70$, $M = 2$, and $G = 1$), subject to $0 \leq G \leq M$. This variation generates differences in objective likelihoods and Bayesian posterior probabilities across tasks. Task order is randomized, and each task is independent.

Provided likelihoods and analysis sample. All participants were shown ordered signal sequences. In the L*/P treatment, however, the likelihood values supplied for mixed-color samples were calculated for the corresponding unordered compositions rather than the exact displayed sequences. Although the likelihood ratio and implied Bayesian posterior remain unchanged, the supplied values overstate the probability of the exact displayed sequence under each bag by a factor of $\binom{M}{G}$. The main analysis excludes the affected observations in L*/P, retaining approximately 72% of observations in this treatment. Appendix Section A.5 compares treatments within unaffected environments and separately reports results for the affected mixed-sample environments.

3.4 Recruitment, Procedures, and Incentives

Participants ($N = 800$) were recruited on Prolific in January 2026.[4] Eligibility was restricted to U.S. residents aged 18 or older with a Prolific approval rating of at least 99%. Appendix

[4]The experiment was pre-registered at the AEA RCT Registry (AEARCTR-0017734), and received IRB exemption from Cornell University's Office of Research Integrity and Assurance under protocol number 0150154.

Table A.1 reports summary statistics and shows that observable characteristics are balanced across treatments.

Procedures. Participants first read an informed consent form and completed an instructional overview explaining drawing with replacement. The overview introduced a single bag and a sequence of draws to illustrate the sampling process, emphasizing that samples are presented in exact order. Participants then answered two comprehension questions and received feedback. Appendix Figures E.5 and E.6 provide screenshots of the instructional and comprehension check screens.

Participants next received treatment-specific instructions introducing the two-bag environment and relevant probability judgments, along with an explanation of the payment scheme and bonus incentives. An additional comprehension check verified understanding before proceeding to the main task. Appendix Figures E.7-E.10 display the two probability-task instructions, comprehension check, and feedback screens from the L/P treatment.

After completing all 15 tasks, participants answered a short open-ended question describing their general approach to the tasks in the experiment. Appendix Figure E.11 provides an example of this screen from the L/P treatment.

Incentives. All participants received a fixed payment of $4 for completing the experiment. In addition, one in five participants was randomly selected for a $10 performance bonus. For selected participants, one of the 15 tasks was randomly chosen for the bonus. In P and L^*/P, where a single posterior belief was reported in each task, that report determined payment. In L/P and L, where multiple probability reports were made within a task (likelihoods for each bag and, in L/P, additionally a posterior), one report was randomly selected for payment. The bonus was awarded if the selected report was within five percentage points of the statistically correct benchmark.[5]

On average, participants took 16 minutes and 24 seconds to complete the experiment, and the average bonus payment for selected participants was $4.69.

4 Experiment 1: Results

This section presents the results of the main experiment. I first document baseline posterior updating, then examine distortions in likelihood judgments and how these propagate to

[5]This incentive structure was chosen over standard scoring rules such as quadratic or binary rules, which can distort reported beliefs. In contrast, threshold-based incentives such as the one used here can promote more truthful reporting by reducing cognitive demands (Danz et al., 2022).

posterior beliefs. I next assess whether eliciting or providing likelihoods improves posterior updating. Finally, I examine how likelihood distortions and posterior under-inference relate across different signal environments, motivating a novel underlying channel for the subsequent experiment.

4.1 Baseline Posterior Updating

As a first check, I examine whether posterior updating in the baseline P treatment reproduces findings from prior studies that elicit posteriors directly. To facilitate comparison with these studies, I present beliefs in both probability levels (expressed on a 0–100 scale) and in log-odds. Bayesian posterior log-odds are given by

$$\log\left(\frac{P(A \mid G)}{P(B \mid G)}\right), \tag{8}$$

with the same transformation applied to reported posteriors.

Figure 3 plots reported posteriors on the vertical axis against their Bayesian benchmarks on the horizontal axis. Panels (a) and (b) show the baseline P treatment in levels and log-odds, respectively. Panels (c) and (d) show the corresponding relationships using task-level mean or median responses compiled across studies in the meta-analysis of Benjamin (2019). Reassuringly, the two samples display similar patterns in both representations: reported beliefs move in the direction favored by the evidence but lie below the identity line when the Bayesian benchmark exceeds the prior and above it when the benchmark falls below the prior. Reports are therefore closer to the prior—50% in levels and zero in log-odds—than Bayes' rule implies, implying under-inference to information.

Panels (a) and (b) also show that the relationship between reported and Bayesian posteriors is more closely approximated by a linear fit in levels than in log-odds in the baseline P treatment, motivating the linear specifications used throughout the paper. Accordingly, I use the following linear specification to quantify deviations from Bayesian updating in the P treatment:

$$\pi_{it} - 50 = \alpha + \beta \times (P_{it} - 50) + \varepsilon_{it}, \tag{9}$$

where π_{it} denotes participant i's reported posterior for Bag A in round t, and P_{it} is the corresponding Bayesian posterior as defined in (3). Centering both variables at 50 isolates updating relative to neutral priors. The slope coefficient β measures the extent to which reported posteriors respond to variation in the Bayesian benchmark, while α captures level bias under neutral evidence.

Table 1 reports the regression results. The estimated slope of 0.689 (SE = 0.016) indicates

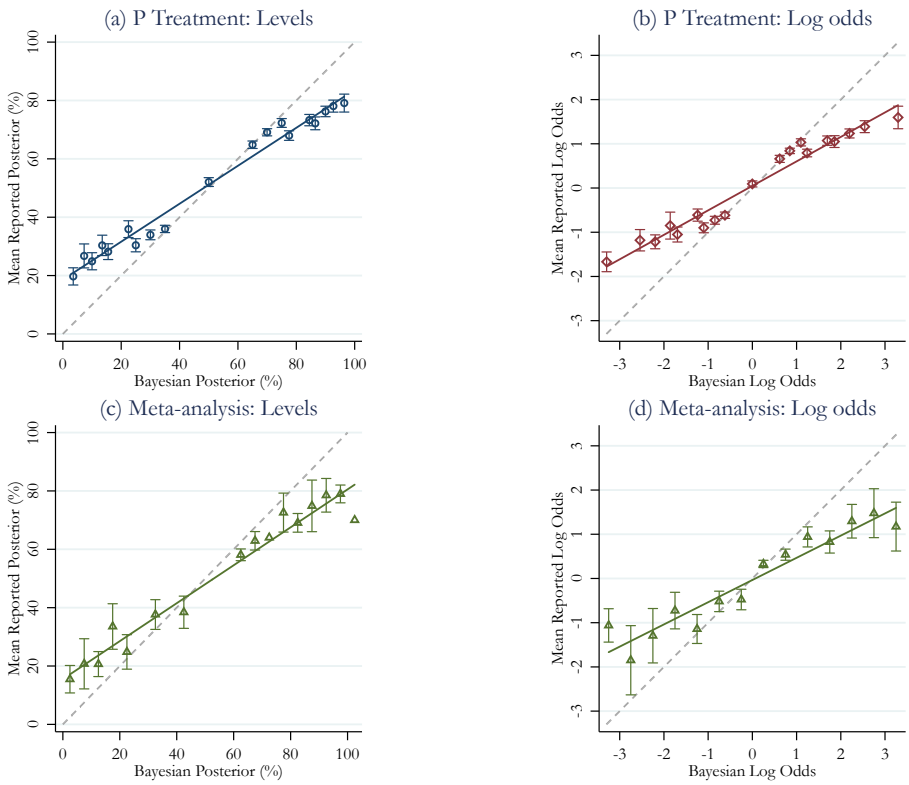


Figure 3. Reported versus Bayesian Posteriors, P Treatment and Meta-Analysis

Notes: Panels (a)–(b) show posterior beliefs in the baseline P treatment in both levels and log-odds, respectively; Panels (c)–(d) show the corresponding relationships using median task-level responses from the meta-analysis of Benjamin (2019). The horizontal axis shows the Bayesian posterior benchmark, and the vertical axis shows the corresponding reported posterior, expressed in the units indicated by each panel. Points denote mean responses within each Bayesian benchmark, and error bars represent 95% confidence intervals. The dashed line denotes the identity line, and solid lines show linear fits.

substantial under-inference relative to the Bayesian benchmark: on average, participants revise their beliefs by only 0.69 percentage points for a one-percentage-point increase in the Bayesian posterior. This estimate is close to the degree of under-inference documented in prior work: re-estimating Equation (9) on the subset of symmetric, equal-prior environments in the meta-analysis of Benjamin (2019) yields a similar slope of 0.653.[6] Finally, although the estimated constant is statistically significant, its magnitude implies only a modest upward bias in posterior beliefs: when the Bayesian posterior equals 50%, the average reported posterior is approximately 51.1%.

Overall, the baseline treatment closely replicates the posterior under-inference documented

[6]The meta-analysis combines task-level mean or median responses, whereas the present regression uses individual reports. Differences in aggregation, weighting, and the composition of updating tasks make this a descriptive comparison rather than an exact replication.

Table 1. Reported Posteriors versus Bayesian Posteriors, P Treatment

	(1)
Bayesian posterior (centered)	0.689***
	(0.016)
Constant	1.121***
	(0.336)
N	3105
R^2	0.66

Notes: The dependent variable is the participant's reported posterior centered at 50; the independent variable is the corresponding Bayesian posterior centered at 50. The slope coefficient measures responsiveness to the Bayesian benchmark, with a value of 1 corresponding to perfect Bayesian updating. The unit of observation is participant–round. Standard errors clustered at the participant level are reported in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in prior work. In the next subsection, I turn to participants' likelihood judgments to examine patterns in the first stage.

4.2 Distortions in Evidence Evaluation

The baseline results show that posterior beliefs respond insufficiently to variation in the Bayesian benchmark, consistent with past work. I now examine whether this arises in evidence evaluation by comparing participants' reported likelihood judgments with the objective likelihoods. Because likelihoods are elicited in both the L/P and L treatments, I also compare these treatments to assess whether eliciting posterior beliefs affects likelihood judgments.

Figure 4 plots mean reported likelihood judgments against the objective likelihoods in the L/P and L treatments, pooling across Bags A and B. For multiple-draw samples ($M > 1$), mean reports lie systematically above the identity line, revealing substantial overestimation in both treatments, with larger errors at lower objective likelihoods. Thus, distortions already arise during evaluating evidence, both in isolation and when posteriors are reported. By contrast, mean reports for single-draw samples ($M = 1$) lie near the identity line. Since these likelihoods correspond directly to the bag compositions, this provides a reassuring check that participants understood the basic reporting task.

To quantify these distortions more formally, I estimate the following specification separately by treatment $g \in \{\text{L/P}, \text{L}\}$:

$$l_{it} = \alpha_g + \beta_g \times L_{it} + \varepsilon_{it}, \tag{10}$$

where l_{it} denotes participant i's reported likelihood in round t, pooled across Bags A and B, and L_{it} is the corresponding objective likelihood as defined in (7). The intercept α_g

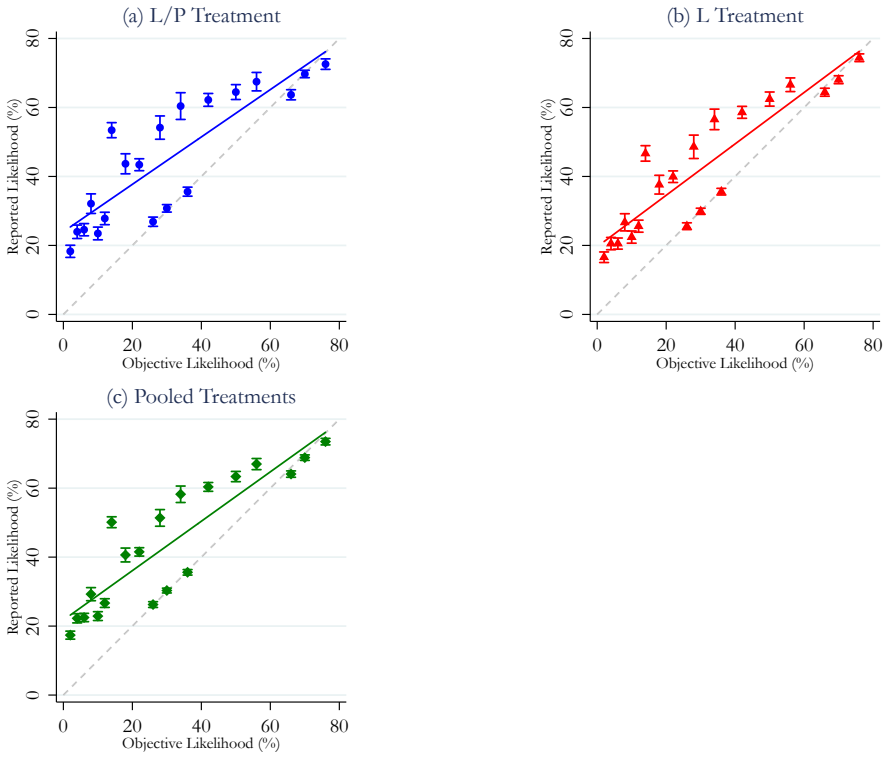


Figure 4. Reported versus Objective Likelihoods by Treatment

Notes: Panels (a)–(c) show reported likelihoods against objective likelihoods in the L/P and L treatments, pooling across Bags A and B; panel (c) shows reported likelihoods pooled across both treatments. Objective likelihoods are binned into 2-percentage-point intervals. Points denote bin means, error bars represent 95% confidence intervals, the dashed line denotes the 45-degree identity line, and solid lines show linear fits.

captures the additive component of the fitted relationship, while β_g measures responsiveness to objective likelihoods. The objective benchmark corresponds to $\alpha_g = 0$ and $\beta_g = 1$; when the slope is close to one, a positive intercept describes an approximately parallel upward shift from the objective benchmark, which quantifies the upward displacement visible in Figure 4. I estimate this specification for multiple-draw samples ($M > 1$), where evaluating likelihoods requires combining the probabilities of successive draws. Appendix Table A.3 reports a more flexible specification that includes single draws and allows both intercepts and slopes to vary with sample size.

Table 2 reports the resulting estimates. Columns (1) and (2) present results separately for the L/P and L treatments, while Column (3) pools observations across both likelihood-elicitation treatments. Across all three columns, the estimates confirm the graphical pattern: large positive intercepts coexist with slopes close to one, indicating that reported likelihoods are systematically distorted upward while remaining responsive to their objective values. The

pooled intercept is 21.3 percentage points, while the treatment-specific intercepts are 23.6 in L/P and 19.1 in L, a statistically significant difference of 4.5 percentage points ($p = 0.002$) when posterior beliefs are also elicited. The estimated slopes range from 0.921 to 0.940, remaining close to one. The difference in slope coefficients across treatments is small and statistically insignificant ($p = 0.60$), suggesting that eliciting posterior beliefs shifts reported likelihoods upward but does not materially affect responsiveness to objective variation. These conclusions are robust even after restricting the sample under pre-registered filters (see Appendix Table A.7).

Table 2. Reported Likelihoods versus Objective Likelihoods by Treatment

	(1) L/P Treatment	(2) L Treatment	(3) Pooled
Objective likelihood	0.921***	0.940***	0.928***
	(0.027)	(0.025)	(0.018)
Constant	23.608***	19.132***	21.317***
	(1.015)	(1.055)	(0.744)
N	3668	4026	7694
R^2	0.34	0.40	0.37

Notes: Columns report linear regressions of reported likelihoods on the corresponding objective likelihood. Columns (1) and (2) estimate the specification separately by treatment, while column (3) pools observations across both likelihood-elicitation treatments. Observations are pooled across Bags A and B. In L/P, participants reported likelihoods and posterior beliefs; in L, participants reported likelihoods only. The unit of observation is participant–round–bag. Standard errors clustered at the participant level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

While these results do not identify the source of these distortions, two patterns suggest that difficulties in compounding probabilities may be a contributing factor. First, Appendix Figure A.1 shows that mean reports closely track objective likelihoods for single signals, when likelihoods correspond directly to bag compositions, but become increasingly distorted for multiple signals. Appendix Table A.3 quantifies this by revealing small intercepts and slopes close to one for single draws, but substantially higher intercepts for multiple draws even though slopes are close to one. Second, open-ended responses from participants in L and L/P mention the use of simplifying multiplicative heuristics such as averaging and guessing, especially in multi-signal environments (see Appendix Section B.4). Although not identical, these patterns are consistent with similar difficulties in evaluating compound event probabilities (e.g., Bar-Hillel, 1973; Tversky and Kahneman, 1983).

To summarize, these results reveal the presence of systematic biases during evidence evaluation, before posteriors are formed. Systematic upward distortions appear both when participants report alone and when participants subsequently report posteriors, indicating

that they do not depend on a requirement to form posterior beliefs. I next examine how these distortions in likelihood judgments translate into posterior beliefs.

4.3 Tracing Posterior Biases to Evidence Evaluation

The preceding analyses document two results so far: posterior beliefs exhibit systematic underinference, and likelihood judgments are systematically distorted. However, this does not show how the two are connected. Participants may correctly combine these distorted likelihoods using Bayes' rule to posterior beliefs, or amplify these distortions when forming posteriors.

To assess how much posterior bias is already encoded in likelihood judgments, I compare reported posteriors with those obtained by applying Bayes' rule to reported likelihoods (Definition 5) in the L and L/P treatments. This comparison reveals whether the second stage of posterior formation adds to or offsets the distortions from evidence evaluation. I examine this mapping within participants in L/P and across independent groups in L and P.

Conditional Bayesian coherence within participants. First, I assess whether participants are conditionally Bayesian given their own reported likelihood judgments. In the L/P treatment, participants report both likelihood judgments and posterior beliefs every round. Under equal priors, the likelihood-implied posterior is therefore

$$\pi^{\ell}_{it} = \frac{\ell_{1it}}{\ell_{1it} + \ell_{2it}}, \tag{11}$$

where ℓ_{bit} is the reported likelihood for bag $b \in \{A, B\}$ by participant i in round t.

Figure 5 plots the mean reported posterior against the corresponding implied posterior obtained from Equation (11) for each signal environment (M, δ, G). Remarkably, the observations lie closely on the identity line across the full space, with the fitted relationship nearly coinciding with the identity benchmark. Appendix Figure A.3 connects this correspondence directly to posterior under-inference by expressing both reported and likelihood-implied posteriors as deviations from the objective Bayesian benchmark, where the two series closely track one another in direction and magnitude. Both results indicate that much of the systematic posterior under-inference is already present in participants' likelihood judgments, with comparatively little additional attenuation during posterior formation.

To quantify this formally, I estimate

$$\pi_{it} - 50 = \alpha + \beta \times \left(\pi^{\ell}_{it} - 50\right) + \varepsilon_{it},$$

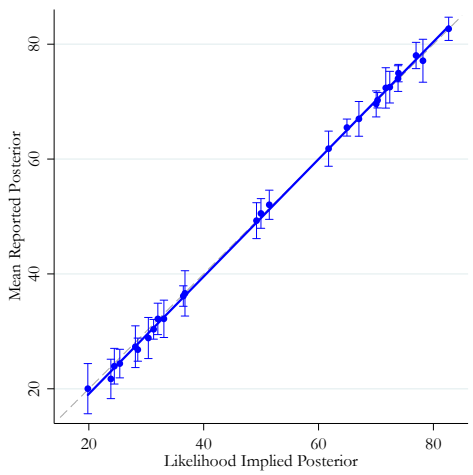


Figure 5. Reported versus Likelihood-Implied Posteriors, L/P Treatment

Notes: Each point corresponds to a signal environment (M, δ, G), which uniquely determines a Bayesian posterior. The horizontal axis plots the posterior implied by participants' reported likelihoods in the L/P treatment, while the vertical axis plots the corresponding mean reported posterior. The dashed line denotes the 45-degree identity line; the solid line shows the linear fit.

where π_{it} is participant i's reported posterior in round t, and π_{it}^{ℓ} is the likelihood-implied posterior constructed Equation (11). Under conditionally Bayesian posterior formation, reported and likelihood-implied posteriors are equivalent, implying $\beta = 1$. A slope below one indicates additional attenuation during posterior formation; a slope above one indicates amplification. I focus on multiple-draw observations ($M > 1$), where systematic likelihood distortions emerge and the objective likelihoods differ from the Bayesian posterior.[7]

Table 3 reports an estimated slope of 0.966 (SE = 0.012), close to the conditional Bayesian benchmark of one. The relationship also has substantial explanatory power at the participant–round level ($R^2 = 0.885$), showing that the correspondence extends beyond alignment between environment-level averages. Reported beliefs therefore move nearly perfectly with the posteriors obtained by applying Bayes' rule to participants' own likelihood judgments.

Overall, participants in L/P report posterior beliefs that are remarkably consistent with their reported likelihoods, even for multiple-draw samples where objective likelihoods no longer coincide with posterior beliefs. Open-ended responses suggest that some participants constructed posteriors explicitly by normalizing their likelihood judgments, highlighting the use of Bayes' rule (Appendix Section B.4). Because both judgments are elicited on the same screen, however, the observed alignment could also reflect efforts to construct mutually consistent reports, which this within-treatment analysis cannot distinguish. Therefore, I

[7]For single draws under the symmetric bag structure and equal priors, the objective likelihood under Bag A coincides with its Bayesian posterior and may reflect this alignment mechanically.

Table 3. Reported Posteriors versus Likelihood-Implied Posteriors in L/P

	(1) L/P Treatment
Likelihood-implied posterior (centered)	0.966***
	(0.012)
Constant	-0.197
	(0.200)
N	1834
R^2	0.885

Notes: The dependent variable is the participant's reported posterior belief centered at 50. The regressor is the posterior implied by the participant's own reported likelihoods, also centered at 50. The sample is restricted to the L/P treatment and to multi-signal observations ($M > 1$). Conditional Bayesian coherence implies an intercept of zero and a slope of one. Standard errors are clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

next turn to independently elicited likelihood judgments to examine whether they reproduce posterior beliefs across treatments where the two are reported separately.

Reproducing baseline posterior beliefs across treatments. I next examine whether independently reported likelihood judgments reproduce baseline posterior beliefs across treatments. To do this, I use Equation (11) and apply Bayes' rule to reports from L, where participants report only likelihoods, and compare the implied posteriors with beliefs in P, where participants report only posteriors. This comparison tests whether the connection between likelihood distortions and posterior under-inference extends beyond settings in which participants report both judgments.

Figure 6 plots the mean reported posterior in P against the corresponding implied posterior in L for each signal environment (M, δ, G). Despite being elicited from independent groups of participants, the two measures again lie closely on the identity line across environments, and the fitted relationship nearly coincides with the identity benchmark. Independently elicited likelihood judgments therefore also reproduce both the level and cross-environment variation in baseline posterior beliefs, showing that the earlier within-participant alignment is not driven solely by jointly eliciting the two reports.

To quantify this relationship formally, I estimate

$$\pi_e^P - 50 = \alpha + \beta \times \left(\pi_e^L - 50\right) + \varepsilon_e,$$

where π_e^P denotes the mean reported posterior in P and π_e^L the posterior implied by likelihood reports in L for signal environment e. Because implied and reported posteriors are not observed for the same individuals, I conduct the analysis at the environment level, focusing

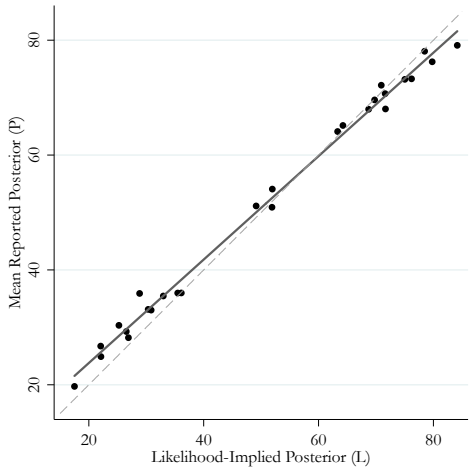


Figure 6. Reported Posteriors in P versus Likelihood-Implied posteriors in L

Notes: Each point corresponds to a signal environment (M, δ, G), which uniquely determines a Bayesian posterior. The horizontal axis plots the posterior implied by mean reported likelihoods in the L treatment, while the vertical axis plots the corresponding mean reported posterior in the P treatment. The dashed line denotes the 45-degree identity line; the solid line shows linear fit.

on $M > 1$ environments.

Table 4 reports the results. The estimated slope of 0.893 (SE = 0.018) indicates a close relationship: a one-percentage-point increase in the mean likelihood-implied posterior in L corresponds to a 0.893-percentage-point increase in the mean reported posterior in P. The estimated intercept is small, indicating little systematic level difference between the two treatment groups.

This tight correspondence holding even across treatments suggests that the alignment in Figure 5 for L/P is not merely an artifact of reporting likelihoods and posteriors together. Participants in L report only likelihoods, yet applying Bayes' rule to these judgments reproduces much of the posterior pattern in P, where likelihoods are never elicited. Independently measured likelihood distortions therefore recover much of baseline posterior under-inference, showing that evidence evaluation is an important source of posterior bias even when participants are asked only to report their final beliefs.

Cross-stage alignment across participants. Finally, I examine whether the aggregate connection between likelihood and posterior responses is also present within individuals. Within L/P, I estimate participant-specific responsiveness to the objective benchmark separately for likelihood judgments and posterior beliefs. The two coefficients are strongly positively related: participants whose likelihood judgments respond more strongly to objective likelihoods also exhibit greater posterior responsiveness, with an estimated coefficient of

Table 4. Reported Posteriors in P versus Likelihood-Implied Posteriors in L

	(1) L vs P Treatments
Likelihood-implied posterior from L (centered)	0.893***
	(0.018)
Constant	0.825*
	(0.404)
N	21
R^2	0.992

Notes: The dependent variable is the mean reported posterior in the P treatment centered at 50. The regressor is the mean posterior implied by likelihood reports in the L treatment, also centered at 50. Each observation corresponds to a signal environment (M, δ, G). The sample is restricted to multi-signal observations ($M > 1$). Conditional Bayesian coherence implies an intercept of zero and a slope of one. Standard errors are heteroskedasticity-robust. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

0.784 (SE = 0.046), suggesting that individual differences in posterior updating are systematically related to differences in evidence evaluation (see Appendix Section B.1 for details of participant-level slopes and intercepts). These patterns are also stable over the experiment; posterior accuracy and response times show little systematic change across rounds within participants, providing little evidence that repeated task exposure substantially improves accuracy (see Appendix Section B.2 and Figure B.3).

Overall, the within-participant, cross-treatment, and individual-level results point to evidence evaluation as an important source of posterior deviations. Reported posteriors closely track those implied by participants' own likelihood judgments, independently elicited likelihoods reproduce most of the baseline underinference, and responsiveness is systematically aligned across stages within individuals. These findings trace a substantial share of posterior deviations to distorted likelihood inputs. The next subsection looks at whether intervening at the evidence-evaluation stage improves posterior beliefs.

4.4 Intervening in Evidence Evaluation

The previous subsection traces much of the observed posterior underinference to distortions in evidence evaluation. I now examine whether intervening at this stage improves posterior beliefs. I compare the baseline P treatment with L/P, in which participants report both likelihood judgments and posteriors, and with L*/P, in which they receive the objective likelihoods and report posteriors. These comparisons identify the effects of two interventions at the evidence-evaluation stage: eliciting participants' own likelihood judgments and supplying

the objective likelihoods directly.[8]

Figure 7 plots mean reported posterior beliefs against the Bayesian benchmark in levels. Panels (a) and (b) display the L/P and L*/P treatments, while Panel (c) reproduces the baseline P treatment from Figure 3 for comparison. Both interventions retain the broad pattern of under-inference observed in the baseline: mean reports lie close to the identity line at moderate Bayesian posteriors but deviate toward 50% at more extreme benchmarks, lying above (below) the identity line below (above) 50%. Despite this, the fitted relationships appear steeper in both intervention treatments than in P, suggesting that eliciting or providing likelihoods improves posterior responsiveness. Appendix Table A.4 reports the corresponding means for each treatment group at the signal environment level.

To quantify the visual patterns in Figure 7 and examine whether likelihood interventions affect posterior beliefs, I re-estimate Equation (9) separately for the L/P and L*/P treatments:

$$\pi_{it} - 50 = \alpha_g + \beta_g \times (P_{it} - 50) + \varepsilon_{it}, \tag{12}$$

where $g \in \{\text{L/P}, \text{L}^*/\text{P}\}$. As before, the slope coefficient β_g measures the responsiveness of reported posterior beliefs to the Bayesian benchmark for each treatment.

Table 5 reports the regression estimates, with Columns (1)–(2) presenting the estimates for the L/P and L*/P treatments, Column (3) reproducing the baseline P treatment for comparison, and Column (4) pooling observations across all posterior-elicitation treatments. Two patterns emerge. First, both interventions improve posterior responsiveness: the estimated slope rises from 0.689 in P to 0.750 in L/P (SE = 0.019) and 0.746 in L*/P (SE = 0.019). The increases are statistically significant ($p = 0.014$ and $p = 0.020$, respectively), representing approximately 9% and 8% greater responsiveness and closing roughly 18–20% of the baseline gap to the unit-slope benchmark.

Second, the estimated improvements are similar across the two interventions: the slope difference between L/P and L*/P is small and statistically insignificant ($p = 0.885$). Therefore, requiring participants to engage with the likelihood stage appears to account for most of the improvement in posterior updating, regardless of whether likelihoods are computed or provided. While intercept differences are statistically significant for L*/P relative to P and L/P, they are small in magnitude: fitted reports at a Bayesian posterior of 50% range from 49.3% to 51.1% across treatments.

These results jointly show that making the evidence evaluation stage explicit improves posterior responsiveness, with similar estimated gains across interventions regardless of

[8]As described in the experimental design, analyses of the L*/P treatment are restricted to observations for which the provided likelihoods correspond to the displayed task. Appendix A.5 compares treatments within unaffected environments and separately reports results for the affected mixed-sample environments.

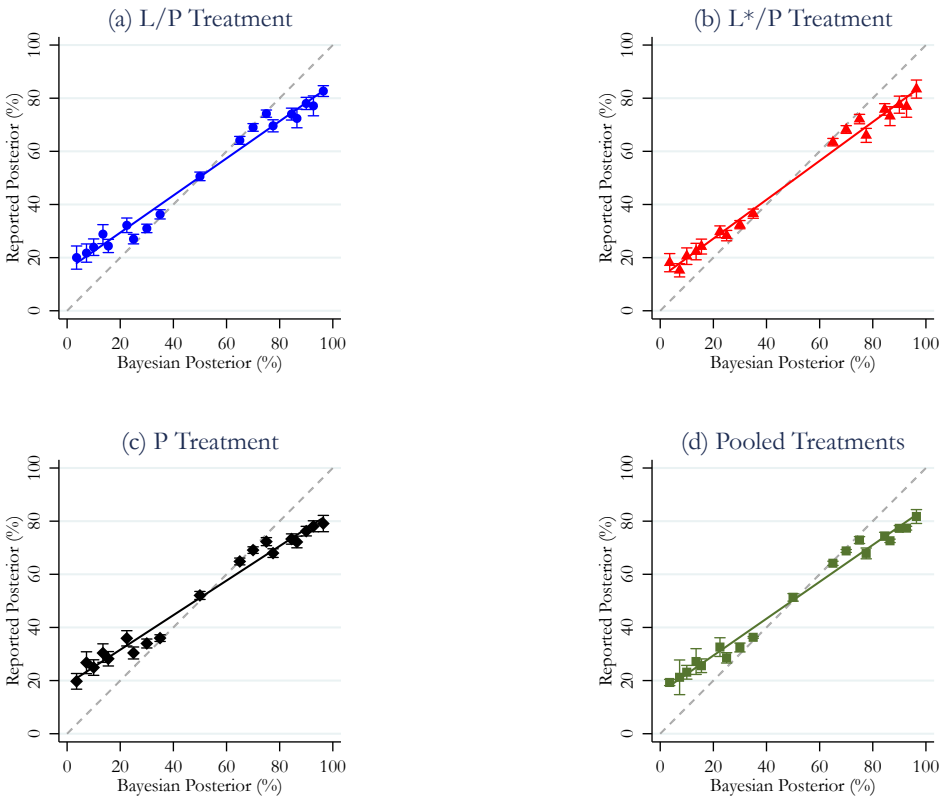


Figure 7. Reported versus Bayesian Posterior Beliefs by Treatment

Notes: Panels (a)–(c) correspond to the L/P, L*/P, and P treatments respectively; panel (d) pools across all three treatments. Each panel plots mean reported posteriors against the corresponding Bayesian posterior. The dashed line denotes the 45-degree identity line. Solid lines show linear fits. Error bars represent 95% confidence intervals.

whether participants compute or are given likelihood judgments. The findings are robust to the pre-registered sample restrictions (Appendix Table A.8) and a specification allowing slopes to vary with the number of signals (Appendix Table A.5).

What drives posterior improvements? Reporting likelihoods could either improve evidence evaluation or help participants translate their judgments into posteriors. In L/P, the results provide little indication that the first channel drives the improvement. Table 2 shows that likelihood judgments in this group are systematically distorted, but reported posteriors closely track those implied by these distorted judgments (Table 3). Open-ended responses from participants in L/P frequently describe multiplying probabilities and normalizing likelihoods, whereas those in P predominantly describe sample proportions or visual

Table 5. Reported Posteriors versus Bayesian Posteriors by Treatment

	(1) L/P Treatment	(2) L*/P Treatment	(3) P Treatment	(4) Pooled
Bayesian posterior	0.750***	0.746***	0.689***	0.727***
	(0.019)	(0.019)	(0.016)	(0.010)
Constant	0.498	-0.656	1.121***	0.415*
	(0.424)	(0.399)	(0.336)	(0.224)
N	2805	2203	3105	8113
R^2	0.69	0.74	0.66	0.70

Notes: Each column reports a linear regression of reported posterior beliefs on the corresponding Bayesian posterior. Columns (1)–(3) estimate the specification separately by treatment group, while Column (4) pools observations across all treatments in which posterior beliefs are elicited. The dependent variable is the participant's reported posterior centered at 50, and the independent variable is the Bayesian posterior centered at 50. The slope coefficient measures responsiveness to the Bayesian benchmark; a value of 1 corresponds to perfect Bayesian updating. The unit of observation is participant–round. Standard errors clustered at the participant level are reported in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

comparisons (Appendix Section B.4). These patterns suggest that reporting likelihoods may help participants use their evaluated evidence more systematically, even when that evaluation remains biased, resulting in improved posteriors.

While providing objective likelihoods in L*/P yields similar improvements in posterior responsiveness, it does not eliminate under-inference entirely. Appendix Figure A.4 makes the contrast clear: reported and likelihood-implied errors closely align in L/P, whereas supplied likelihoods imply zero error in L*/P but reported posteriors remain attenuated. While the results cannot identify why providing correct likelihoods does not close this gap completely, there are two suggestive explanations. First, participants may incorporate likelihoods more fully when expending mental effort to construct them, as in L/P, rather than receiving the values directly. Second, as described at the end of Section 3.3, the supplied likelihoods were incorrect for 27.6% of L*/P observations. Appendix Table A.6 shows greater responsiveness in L*/P than in P within unaffected environments, but no corresponding improvement within affected environments. Although the main analysis excludes affected observations, exposure to them may have influenced responses in other rounds, which cannot be ruled out.

4.5 Likelihood Distortions and Relative Evidence

The preceding analyses link distorted likelihood judgments to posterior under-inference, but this relationship varies across signal environments. Substantial likelihood overestimation accompanies pronounced under-inference in some environments and relatively accurate posterior beliefs in others, seemingly creating a puzzle. For example, mean posterior reports

are close to the Bayesian benchmark for both single signals and three-signal samples containing both colors, even though likelihood judgments are substantially distorted in the latter (see Appendix Tables A.2 and A.4).

To understand this variation, Bayesian updating provides a benchmark for how likelihoods translate into posterior beliefs. Under equal priors, posterior beliefs depend only on the ratio of the two likelihoods (Definition 2), which measures the relative support for the two states. Consider a single green signal ($M = 1, G = 1$) versus a sequence containing two green signals and one orange signal ($M = 3, G = 2$). With $\delta = 0.70$, their objective likelihood pairs are $(0.70, 0.30)$ and $(0.147, 0.063)$, respectively. Although the three-signal sequence has lower objective likelihoods under both bags, the likelihoods fall proportionally, leaving their relative weight unchanged; both therefore result in the same Bayesian posterior of 70%. In contrast, three green signals ($M = 3, G = 3$) yield an objective likelihood pair of $(0.343, 0.027)$. The likelihood for Bag B is proportionally smaller for this sequence, producing stronger relative support for Bag A, which translates to a stronger posterior of approximately 92.7%.

Figure 8 examines how participants' judgments depart from these benchmarks across all sample size and composition pairs (M, G). Panel (a) reports mean likelihood errors, defined as reported minus objective likelihoods, separately for the bag under which the sample is more likely and the bag under which it is less likely. Errors are small for single signals but substantial for multiple signals, with participants overestimating the sample's likelihood under both bags. These upward errors can be especially consequential in proportional terms when an objective likelihood is small. From the above example, the objective likelihood of three green signals under the mostly-orange bag is only 2.7% when $\delta = 0.70$, so even a few percentage points of overestimation would substantially inflate it. Whether this weakens relative evidence, however, depends on how much the likelihood under the other bag is inflated.

I examine this relative distortion directly in Panel (b), which reports the mean fraction of the objective log likelihood ratio retained in participants' judgments, calculated individually as

$$\frac{\log(\ell_H/\ell_L)}{\log(L_H/L_L)}, \tag{13}$$

where ℓ_b and L_b denote reported and objective likelihoods, respectively, and H and L identify the bags with higher and lower objective likelihoods. A value of one indicates preservation of relative evidence, while values between zero and one indicate compression. The panel distinguishes pure samples, containing signals of only one color, from mixed samples, containing signals of both colors. As the preceding example illustrates, pure samples can make the observed sequence particularly unlikely under one bag, whereas signals of different

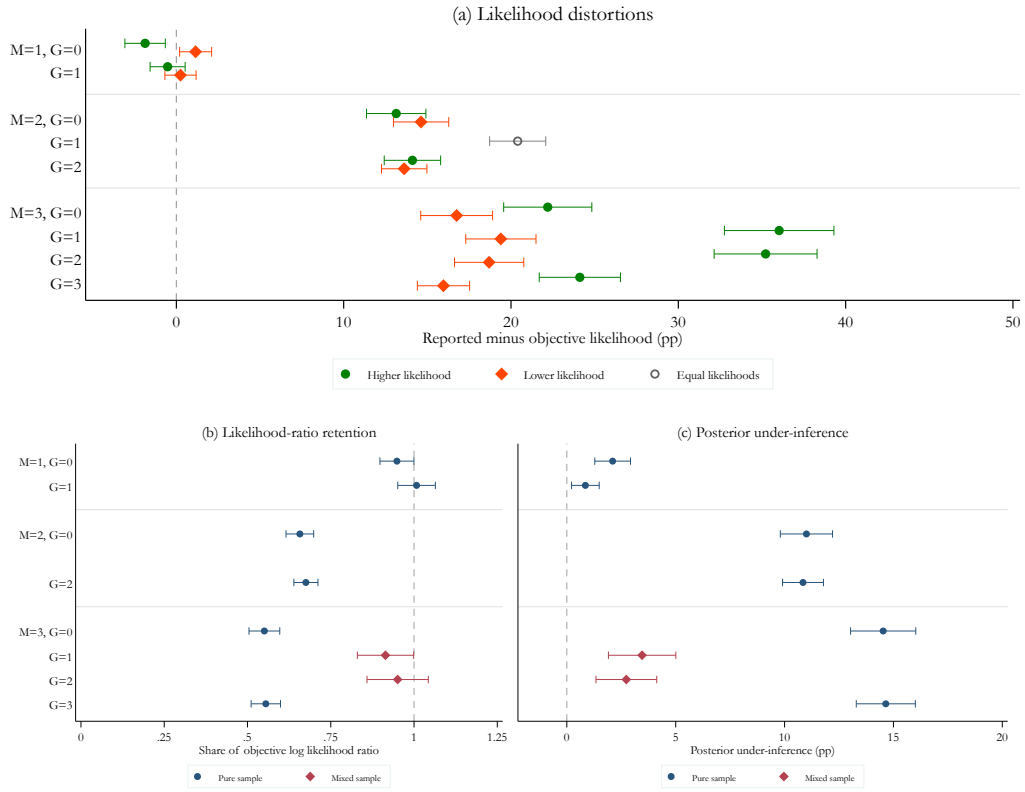


Figure 8. Likelihood and Posterior Distortions Across Signal Environments

Notes: Each row identifies sample size and composition (M, G), pooling across bag compositions. Panels (a)–(b) pool L/P and L. Panel (a) reports mean likelihood errors (reported minus objective), distinguishing bags by their objective likelihoods. Panel (b) reports the corresponding mean log-likelihood-ratio retention from Equation (13), calculated at the individual level. Panel (c) pools P, L/P, and L*/P and reports posterior errors oriented so that positive values indicate under-inference toward 50%. Panel (b) excludes zero likelihood reports; Panels (b)–(c) omit $(M = 2, G = 1)$, where the Bayesian posterior equals the prior. Error bars show 95% confidence intervals clustered by participant.

colors provide offsetting relative evidence.

Panel (b) shows that relative evidence is largely preserved for single signals. Among multiple-signal samples, however, compression is substantially greater for pure samples. For three signals, participants retain approximately 55% of the objective log likelihood ratio in pure samples, compared with more than 90% in mixed samples. Thus, the large percentage-point likelihood distortions in mixed three-signal samples from Panel (a) largely preserve relative support. In pure samples, by contrast, compression indicates that the lower likelihood is disproportionately inflated relative to the higher likelihood. This distinction connects the ratio results to the underlying likelihood levels: when the lower objective likelihood is particularly small, even an error smaller in percentage points can represent a much larger

proportional distortion, compressing relative support.

Panel (c) shows that posterior under-inference mirrors these differences in relative evidence. The measure is the Bayesian posterior minus the reported posterior, with the sign reversed below 50% so that positive values indicate attenuation toward the prior. For three-signal samples, the substantial likelihood-ratio compression in pure samples accompanies approximately 14–15 percentage points of under-inference. In contrast, mixed samples exhibit only around 3 percentage points of under-inference: despite their large likelihood errors in Panel (a), relative support is largely preserved in Panel (b). This indicates that what drives posterior bias is not simply whether participants overestimate likelihoods, but how that overestimation changes the relative weight or balance between the two bags.

This distinction suggests a novel role for *absolute likelihood levels* in explaining posterior bias. The same percentage-point likelihood distortion can have different consequences for relative evidence depending on the objective likelihood to which it applies. If both likelihood judgments contain an additive upward distortion, changing the objective likelihood levels could alter subjective relative support and posteriors even within objective likelihood ratios. Section 5 tests this channel directly through a controlled experiment.

4.6 Discussion

The results from the main experiment examine and connect biases across the two stages of belief updating. Making the intermediate evidence-evaluation stage observable reveals systematic distortions in likelihood judgments, suggesting that deviations from the Bayesian benchmark emerge before posterior beliefs are even formed. Subsequently, a large portion of posterior under-inference is already reflected in these distorted likelihood judgments. Within-treatment, reported posteriors closely track those obtained by applying Bayes' rule to participants' own likelihood reports; across treatments, independently reported likelihood judgments reproduce much of the posterior under-inference observed in the baseline treatment. Overall, these comparisons jointly identify evidence evaluation as an important and systematic source of posterior bias.

The randomized treatment comparisons additionally show that making likelihoods explicit improves posterior responsiveness. Reporting likelihood judgments and being supplied objective likelihoods produce similar estimated improvements, suggesting that explicitly engaging with likelihoods can help participants form more accurate posterior beliefs, even when their evaluation of the evidence remains distorted. The main likelihood and posterior findings are robust under the pre-registered sample restrictions, and the analysis across rounds finds no evidence of learning over time.

Finally, the comparisons across signal environments show that large likelihood distortions do not always correspond with large posterior errors. This suggests that absolute likelihood levels – a channel overlooked by Bayesian updating – may help shed light on this puzzle by examining how the upward likelihood distortions documented alter their relative support. In the main experiment, however, likelihood levels vary alongside relative evidentiary strength and sample composition, which makes it difficult to isolate this channel. The next section separates and examines this channel through a controlled experiment.

5 Experiment 2: Likelihood Levels and Beliefs

The preceding results document systematic upward distortions in likelihood judgments, but show that the effect on posterior beliefs varies by signal environment. This suggests that the underlying *absolute likelihood levels* may be a novel channel to explain how these distortions translate into beliefs. A common additive upward distortion, such as that observed in the main experiment, would compress the likelihood ratio more strongly when both likelihoods are small, generating greater under-inference under conditionally Bayesian posterior formation. Beliefs may therefore vary by likelihood levels even when the relative strength is unchanged, contrary to the predictions of Bayesian updating and a wide class of behavioral models.

The main experiment cannot isolate this channel because changes in sample composition also alter likelihood levels and relative evidentiary strength. Therefore, in a second experiment, I construct matched environments to vary absolute likelihood levels while holding the objective likelihood ratio—and hence the Bayesian posterior—unchanged. I first formalize the testable predictions, then present the experimental design and results.

5.1 Theoretical Framework and Predictions

To derive testable predictions for how likelihood levels affect posterior beliefs, it is useful to distinguish two broad categories of modeling belief departures from Bayesian updating. To illustrate this, I first express Bayes' rule (Definition 3) in odds form by taking the ratio of the two posterior beliefs:

$$\underbrace{\frac{P(A \mid S)}{P(B \mid S)}}_{\text{Posterior odds}} = \underbrace{\frac{L_A(S)}{L_B(S)}}_{\text{Likelihood ratio}} \times \underbrace{\frac{P(A)}{P(B)}}_{\text{Prior odds}} . \tag{14}$$

Traditionally, past work modeling belief biases can be grouped into one of two broad

categories.[9] The first category models the subjective likelihoods themselves, allowing sampling beliefs, representativeness, or attention to distort the objective likelihood inputs, $L_A(S)$ and $L_B(S)$ in Equation (14). These distorted inputs are then combined in an approximately Bayesian fashion (e.g., Rabin, 2002; Gennaioli and Shleifer, 2010; Benjamin et al., 2016).

The second category explains updating errors as individuals incorrectly combining objective likelihood and prior odds by assigning different weights to the likelihood ratio, prior odds, or both (e.g., Edwards, 1968; Grether, 1980; Holt and Smith, 2009; Benjamin et al., 2019). Stable weights can generate underreaction, overreaction, or base-rate neglect, but imply identical posterior beliefs whenever the objective likelihood ratio and priors are unchanged. More generally, this invariance holds for any stable mapping from the objective likelihood ratio and prior odds to posterior beliefs.

This invariance provides the benchmark for examining how absolute likelihood levels affect beliefs. The upward likelihood distortions documented in the main experiment suggest a reason this invariance might fail: distortions in individual likelihood levels can change their subjective ratio even when the objective ratio is held fixed. A simple illustration below shows how.

An illustrative mechanism. To show how upward likelihood distortions can generate posterior under-inference, suppose $\widetilde{L}_b = \alpha + \beta L_b$ for $b \in \{A, B\}$, with common parameters $\alpha > 0$ and $\beta > 0$ restricted so that subjective likelihoods remain valid probabilities over the environments considered. This specification is motivated by the experimental results in Section 4.2. The subjective likelihood ratio can thus be expressed as

$$\widetilde{R} = \frac{\alpha + \beta L_A}{\alpha + \beta L_B}.$$

The common additive component draws this ratio toward one. Its influence depends on the absolute likelihood levels: scaling both objective likelihoods upward reduces the relative importance of α, moving $\widetilde{R}$ toward the unchanged objective ratio $R = L_A/L_B$. Under equal priors and conditionally Bayesian posterior formation, this compression implies

$$\widetilde{P} - \frac{1}{2} = \frac{\beta(L_A + L_B)}{2\alpha + \beta(L_A + L_B)} \left(P - \frac{1}{2} \right), \tag{15}$$

where $P = L_A/(L_A + L_B)$ is the objective Bayesian posterior and $\widetilde{P}$ is the posterior implied by subjective likelihoods. The coefficient multiplying $P - 1/2$ lies between zero and one, so the additive distortion attenuates posterior beliefs toward the prior, with less attenuation at

[9] While these approaches need not be mutually exclusive, they direct attention to different sources of posterior error.

higher likelihood levels for a fixed objective ratio. A purely proportional distortion ($\alpha = 0$), in contrast, would leave relative evidence and the implied posterior unchanged (see Section D.1 for derivation).

Likelihood-ratio sufficiency. The affine example illustrates one mechanism through which absolute likelihood levels may matter, but the experimental test below does not assume that it describes the underlying evaluation process. I instead formulate a restriction allowing any stable transformation of the objective likelihood ratio. Let

$$R(E) = \frac{L_A(E)}{L_B(E)}, \qquad \widetilde{R}(E) = \frac{\widetilde{L}_A(E)}{\widetilde{L}_B(E)},$$

where $\widetilde{R}(E)$ measures relative evidence constructed from participants' elicited likelihood judgments. The restriction asks whether this subjective ratio depends on the environment only through $R(E)$.

Definition 6 (Likelihood-Ratio Sufficiency)**.** *Evidence evaluation satisfies likelihood-ratio sufficiency if subjective relative evidence depends on the information environment only through its objective likelihood ratio. That is, there exists a function $g(\cdot)$ such that*

$$\widetilde{R} = g(R). \tag{16}$$

This restriction does not require accurate evidence evaluation; it requires only that environments with the same objective likelihood ratio generate the same subjective likelihood ratio. For example, $\widetilde{R} = R^{\gamma}$ compresses relative evidence when $0 < \gamma < 1$ and amplifies it when $\gamma > 1$, while $\gamma = 1$ corresponds to Bayesian updating. Allowing an arbitrary g permits more general distortions but preserves the following invariance.

Proposition 1 (Ratio Invariance)**.** *Suppose evidence evaluation satisfies likelihood-ratio sufficiency. For any two information environments E and E',*

$$R(E) = R(E') \quad \Longrightarrow \quad \widetilde{R}(E) = \widetilde{R}(E').$$

If the environments have the same prior and posterior formation is conditionally Bayesian, then the same condition also implies $\widetilde{P}(A \mid E) = \widetilde{P}(A \mid E')$.

The first experimental comparison tests this restriction by varying absolute likelihood levels at a fixed objective ratio. If subjective ratios differ, no single transformation $g(R)$ can characterize evidence evaluation across these environments, regardless of its flexibility.

Posterior reports separately test whether final beliefs are invariant, as predicted by stable mappings from objective likelihood and prior odds. Observing both stages therefore reveals whether differences in posterior beliefs are accompanied by differences already present in subjective relative evidence.

Likelihood-pair sufficiency. While the first comparison tests whether absolute likelihood levels matter while holding the realized sample fixed, the differences between pure and mixed samples from Section 4.5 suggest that sample composition may be an important input even under identical objective likelihoods, as further documented in past work (e.g., Griffin and Tversky, 1992; Rabin, 2002). Holding both likelihoods fixed allows me to test whether composition affects subjective judgments beyond its implications for the sample's probability under either state, as formulated below.

Definition 7 (Likelihood-Pair Sufficiency)**.** *Evidence evaluation satisfies likelihood-pair sufficiency if there exists a function $F(\cdot,\cdot)$ such that*

$$(\widetilde{L}_A, \widetilde{L}_B) = F(L_A, L_B). \tag{17}$$

This restriction permits distortions that depend flexibly on both objective likelihood levels. It requires only that environments with identical likelihood pairs generate identical subjective likelihood judgments, as formalized below.

Proposition 2 (Sample Invariance)**.** *Suppose evidence evaluation satisfies likelihood-pair sufficiency. For any two environments E and E',*

$$\big(L_A(E), L_B(E)\big) = \big(L_A(E'), L_B(E')\big) \quad \Longrightarrow \quad \big(\widetilde{L}_A(E), \widetilde{L}_B(E)\big) = \big(\widetilde{L}_A(E'), \widetilde{L}_B(E')\big).$$

If the environments have the same prior and posterior formation is conditionally Bayesian, then reported posterior beliefs also coincide.

The second experimental comparison tests this restriction by comparing a pure sample with a mixed one, both with the same number of signals and identical objective likelihood pairs. Differences in reported likelihoods and posteriors would identify a role of sample composition even when its likelihood under each state is unchanged.

Propositions 1 and 2 therefore motivate the experimental design below by examining distinct features of evidence evaluation. The first tests whether subjective relative evidence changes with absolute likelihood levels at a fixed objective ratio and realized sample. The second tests whether subjective likelihood judgments change with sample composition at a

fixed objective likelihood pair. Posterior reports then reveal whether these differences are also reflected in final beliefs.

5.2 Experimental Design

The experimental design uses asymmetric bag compositions to vary how likely the same observed sample is under both states without changing its relative support for either state. I vary the proportions of green and orange balls in each bag independently to construct environments in which both likelihoods change by the same factor. This changes absolute likelihood levels while preserving the likelihood ratio and the Bayesian posterior, isolating the effect of likelihood levels on belief formation. The task and interface otherwise follow the main experiment, except that bag compositions are displayed as percentages rather than counts (see Appendix Figures E.13-E.16).[10]

Figure 9 summarizes the matched environments. Panel A tests ratio invariance (Proposition 1) by varying absolute likelihood levels within three likelihood-ratio conditions across 16 distinct environments. Each comparison is implemented separately for one green signal and two green signals. For example, in the $R = 1.78$ condition, the low- and high-likelihood environments have likelihood pairs $(0.16, 0.09)$ and $(0.64, 0.36)$, respectively across both one and two signals. This proportional scaling makes the same observed sample four times as likely under either bag while preserving the likelihood ratio, Bayesian posterior, and the sample composition.

Panel B tests sample invariance (Proposition 2), motivated by the differences between pure and mixed samples in the main experiment. Whereas Panel A holds composition fixed to isolate likelihood levels, Panel B compares a pure and a mixed two-draw sample with identical objective likelihoods, $(0.16, 0.09)$, and hence the same Bayesian posterior. This comparison tests whether composition affects subjective evidence beyond its implications for the objective likelihood pair. Appendix Table C.1 reports the bag compositions and parameters for each environment.

Treatment Groups. Participants are randomly assigned to the P or L/P treatment. As in the main experiment, participants in P report only posterior beliefs, while participants in L/P report likelihood judgments and posterior beliefs. The L/P treatment makes the evidence-evaluation stage observable, allowing the environmental comparisons to identify effects on subjective likelihoods and the resulting propagation into posterior beliefs. The P

[10]Some differences in bag composition are smaller than in the first experiment. Displaying percentages avoids making visual comparisons of nearly identical 100-ball arrays an additional feature of the task.

Panel A: Likelihood Levels

R	(M, G)	**Low**	**Medium**	**High**
1.78	$(1, 1)$	$(0.16,\ 0.09)$	—	$(0.64,\ 0.36)$
1.78	$(2, 2)$	$(0.16,\ 0.09)$	—	$(0.64,\ 0.36)$
2.25	$(1, 1)$	$(0.09,\ 0.04)$	$(0.36,\ 0.16)$	$(0.81,\ 0.36)$
2.25	$(2, 2)$	$(0.09,\ 0.04)$	$(0.36,\ 0.16)$	$(0.81,\ 0.36)$
4.00	$(1, 1)$	$(0.16,\ 0.04)$	$(0.36,\ 0.09)$	$(0.64,\ 0.16)$
4.00	$(2, 2)$	$(0.16,\ 0.04)$	$(0.36,\ 0.09)$	$(0.64,\ 0.16)$

Panel B: Sample Composition ($M = 2$)

Pure ($G = 2$)	**Mixed ($G = 1$)**
$(0.16,\ 0.09)$	$(0.16,\ 0.09)$

Figure 9. Experimental Environments for Experiment 2

Notes: Cells report objective likelihood pairs (L_A, L_B) for the displayed ordered sequence; M denotes sample size and G the number of green draws. Panel A varies likelihood levels at a fixed sample and likelihood ratio. Its eight configurations are each implemented with one and two draws, yielding 16 environments; dashes indicate omitted configurations. Panel B compares a pure and a mixed sample with identical likelihood pairs. Appendix Table C.1 reports all parameter values.

treatment measures the corresponding posterior effects when likelihood evaluation remains implicit.

Participants and procedures. I recruited 200 U.S. participants on Prolific in August 2026, restricting eligibility to individuals aged 18 or older with a Prolific approval rating of at least 99%.[11] The instructions, comprehension checks, and performance-based incentives were otherwise identical to the main experiment, as detailed in Section 3.4.

Each participant completed 20 tasks presented in randomized order, comprising the 16 distinct Panel A environments and the two Panel B environments. Two randomly selected Panel A environments were additionally repeated for each participant. Appendix Table C.2 reports summary statistics and shows that observable characteristics are balanced across treatments. Participants completed the experiment in an average of 18 minutes and 10 seconds and earned an average of $5.10, including performance-based bonuses.

[11]The experiment was preregistered at the AEA RCT Registry (AEARCTR-0019513) and received an exemption from Cornell University's Office of Research Integrity and Assurance under protocol number 0151094.

5.3 Likelihood Judgments and Likelihood-Ratio Sufficiency

I first examine whether absolute likelihood levels affect reported likelihood judgments and the relative evidence they convey, using participants in L/P where likelihood judgments were reported. The distinction matters because errors in the two likelihoods can preserve their ratio, and even a distorted subjective ratio need not violate likelihood-ratio sufficiency (Definition 6) if it remains unchanged across environments with the same objective ratio.

Figure 10 plots mean likelihood bias—reported minus objective likelihood—separately for Bags A and B across likelihood levels within each objective-ratio condition. Across all three conditions, participants systematically overestimate likelihoods for both bags to a greater extent at low than at high likelihood levels. Thus, likelihood distortions systematically increase when the underlying likelihood levels decrease, despite the objective ratio remaining fixed. Appendix Figure C.1 shows that the same pattern is present in both the one- and two-draw implementations, although overestimation is systematically larger in the latter, consistent with the findings from the main experiment.

To quantify these differences across likelihood levels, I estimate

$$l_{ibe} - L_{be} = \alpha + \beta_1 \times \text{Medium}_e + \beta_2 \times \text{High}_e + \gamma_{r(e)} + \lambda_b + \varepsilon_{ibe}, \tag{18}$$

where l_{ibe} denotes participant i's reported likelihood for bag b in environment e, and L_{be} is the corresponding objective likelihood. The dependent variable is likelihood bias, defined as the difference between the reported and objective likelihood. The regression pools observations from Bags A and B. The omitted category is the low-likelihood environment, with the coefficients β_1 and β_2 measuring changes in likelihood bias relative to the low-likelihood environment. Likelihood-ratio fixed effects, $\gamma_{r(e)}$, compare environments with the same likelihood ratio, while bag fixed effects, λ_b, account for average differences between Bags A and B.

Table 6 reports the results. Estimated bias decreases by 6.64 percentage points in high- compared to low-likelihood environments ($p < 0.001$) within likelihood-ratio conditions and controlling for bag fixed effects. The high–medium difference is also statistically significant ($p < 0.001$), while the medium–low difference is small and statistically insignificant. Thus, the greatest distortions in likelihood bias occur in the low-likelihood environments and decrease systematically as the underlying likelihood level increases.

Whether these differences alter relative evidence depends on how the two reported likelihoods change together, since proportionally similar errors can leave their ratio unchanged. Figure 11 therefore plots the median subjective likelihood ratios—the ratios of the two reported likelihoods—directly across likelihood levels for each objective ratio, separately for

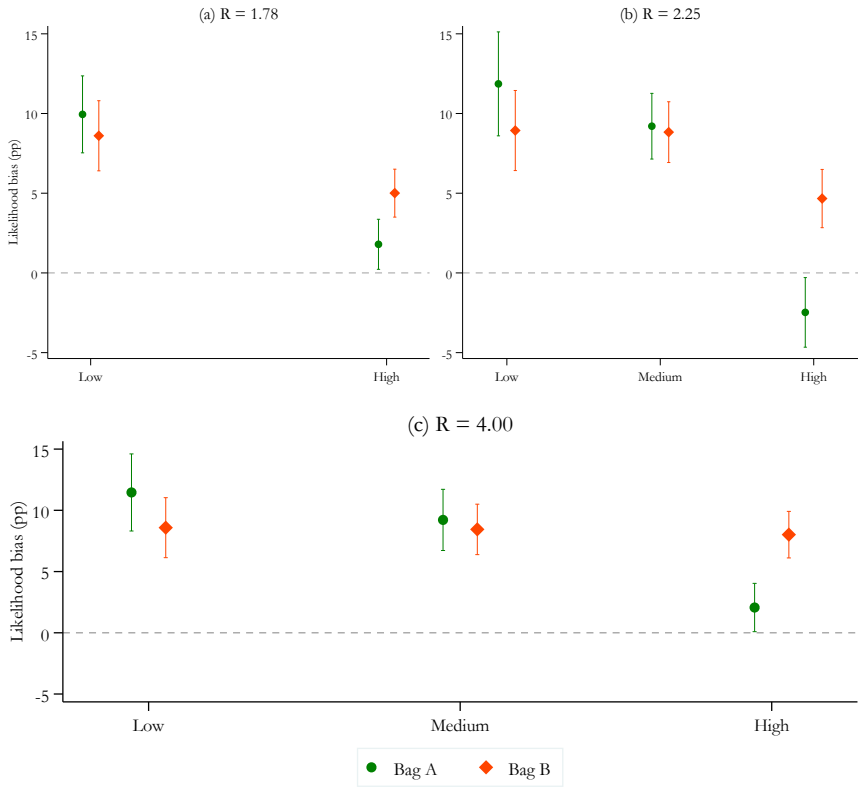


Figure 10. Likelihood Bias by Likelihood Level and Objective Ratio

Notes: Each panel corresponds to one objective likelihood-ratio condition $R = L_A/L_B$ and compares low-, medium- (where available), and high-likelihood environments from Panel A of Figure 9 for participants in L/P. Points show mean reported minus objective likelihoods, in percentage points, separately for Bags A and B. The dashed line indicates zero bias; error bars show 95% confidence intervals.

one- and two-draw samples. For one-draw samples, median subjective ratios closely match the objective benchmark at every likelihood level. For two-draw samples, however, they lie systematically between one and the objective ratio, indicating compression of relative evidence. This compression is greater at lower absolute likelihood levels: within each objective-ratio condition, the median subjective ratio is lower in the low- than in the high-likelihood environment.

These results show that lower absolute likelihood levels generate greater likelihood overestimation and, for two-draw samples, are associated with greater compression of subjective relative evidence. The variation in median subjective ratios suggests a departure from likelihood-ratio sufficiency: environments with the same objective ratio need not convey the same subjective relative support. I next examine whether posterior beliefs also vary across these objectively equivalent environments.

Table 6. Likelihood Bias by Likelihood Level

	(1) Likelihood-reporting bias
Medium likelihood	-0.76
	(0.76)
High likelihood	-6.64***
	(1.31)
Constant	8.90***
	(1.19)
N	3348
R^2	0.048

Notes: The sample consists of participants assigned to the L/P treatment. The dependent variable is likelihood bias, defined as the reported likelihood minus the true likelihood. The omitted category is the low-likelihood environment. The regression pools observations from Bags A and B and includes likelihood-ratio and bag fixed effects. The specification includes likelihood-ratio and bag fixed effects. Standard errors clustered at the participant level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.4 Posterior Beliefs and Ratio Invariance

The preceding analysis shows that absolute likelihood levels affect reported likelihood judgments and alter subjective relative evidence even when the objective ratio is fixed. I next examine whether reported posterior beliefs vary across these matched environments, testing ratio invariance from Proposition 1. Because the objective likelihood ratio and Bayesian posterior remain fixed within each comparison, differences in reported beliefs reveal whether absolute likelihood levels matter beyond objective relative support.

Figure 12 plots posterior bias, defined as the Bayesian posterior minus the reported posterior, within each objective likelihood-ratio condition and separately for the P and L/P treatments. Ratio invariance predicts that bias remains constant across likelihood levels within each ratio. Instead, posterior bias declines systematically with likelihood levels in both treatments, with the greatest underinference in low-likelihood environments and similar patterns across treatments. In the $R = 1.78$ and $R = 2.25$ conditions, the bias changes sign: environments with identical relative evidence and Bayesian posteriors generate underinference at low likelihood levels but overinference at high levels. Thus, changing how likely the same observed sample is under both bags can reverse the direction of posterior error despite no change in relative evidentiary strength.

To quantify these differences formally, I estimate

$$p_e - \pi_{ie} = \alpha + \beta_1 \times \text{Medium}_e + \beta_2 \times \text{High}_e + \gamma_{r(e)} + \varepsilon_{ie}, \tag{19}$$

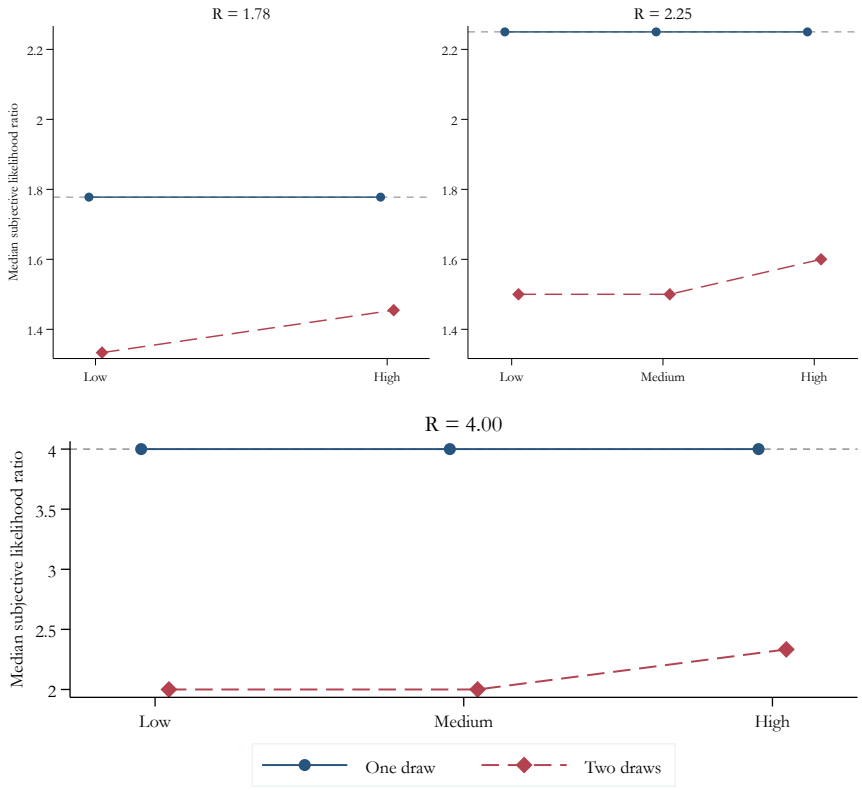


Figure 11. Median Subjective Likelihood Ratios by Likelihood Level and Objective Ratio

Notes: Each panel corresponds to one objective likelihood-ratio condition $R = L_A/L_B$ and compares low-, medium- (where available), and high-likelihood environments from Panel A of Figure 9 for participants in L/P. Points report median subjective likelihood ratios separately for one-draw and two-draw samples across participants; observations with a zero denominator are excluded. Connecting lines link these medians, and horizontal dashed lines mark the objective likelihood ratio.

where e indexes an experimental environment, defined by the bag compositions and observed sample, and $r(e)$ denotes its likelihood-ratio condition. π_{ie} denotes participant i's reported posterior belief, p_e is the corresponding Bayesian posterior, and $\gamma_{r(e)}$ denotes likelihood-ratio fixed effects. The dependent variable is posterior bias, defined as the difference between the Bayesian posterior and the reported posterior, so positive values indicate under-inference. The omitted category is the low-likelihood environment. I estimate Equation (19) separately for the P and L/P treatments, with β_1 and β_2 measuring changes in posterior bias relative to the low-likelihood environment while holding the likelihood ratio fixed.

Table 7 confirms the graphical evidence across both treatments. Column (1) reports the P estimates, where posterior bias is 2.07 percentage points lower in medium-likelihood environments and 7.17 percentage points lower in high-likelihood environments than in low-

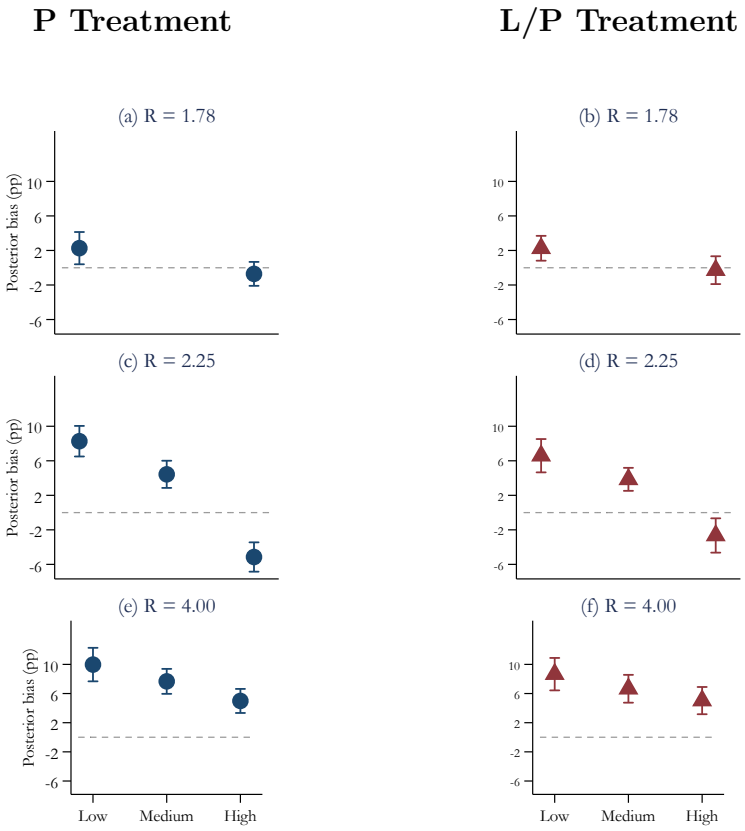


Figure 12. Posterior Bias by Likelihood Level and Objective Ratio

Notes: Panels (a)–(b), (c)–(d), and (e)–(f) correspond to objective likelihood ratio conditions ($R = L_A/L_B$) 1.78, 2.25, and 4.00, respectively. The left column reports the P treatment and the right column the L/P treatment. Eacn panel compares low-, medium- (where available), and high-likelihood environments within one ratio from Panel A of Figure 9. Posterior bias is defined as the Bayesian posterior minus the reported posterior, in percentage points, so positive values indicate under-inference. Points denote mean posterior bias and error bars denote 95% confidence intervals. The dashed horizontal line indicates zero bias.

likelihood environments. Column (2) shows the corresponding differences in L/P, 1.75 and 5.17 percentage points. All four differences are statistically significant ($p < 0.01$), as are the medium–high comparisons in both treatments ($p < 0.001$). Thus, the systematic relationship with absolute likelihood levels persists even when the objective ratio is fixed, regardless of whether participants report posteriors alone or also evaluate likelihoods explicitly.

To connect these posterior responses to the preceding likelihood results, I examine whether participants' likelihood judgments in L/P imply a similar dependence on absolute likelihood levels. Under equal priors, the likelihood-implied posterior, expressed on the same 0–100 scale, is

Table 7. Posterior Bias by Likelihood Level and Treatment

	(1) P	(2) L/P	(3) L/P (Implied)
Medium likelihood	-2.07***	-1.75***	-0.34
	(0.74)	(0.66)	(0.46)
High likelihood	-7.17***	-5.17***	-0.74**
	(0.82)	(0.72)	(0.37)
Constant	4.40***	3.52***	2.45***
	(0.85)	(0.75)	(0.42)
N	1926	1674	1674
R^2	0.108	0.083	0.039

Notes: Columns (1) and (2) define posterior bias as the Bayesian posterior minus the reported posterior for the P and L/P treatments, respectively. Column (3) replaces the reported posterior with the posterior implied by participants' reported likelihood judgments in L/P. The omitted category is the low-likelihood environment. All specifications include likelihood-ratio fixed effects. Standard errors clustered at the participant level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$$\tilde{p}_{ie} = \frac{\ell_{iAe}}{\ell_{iAe} + \ell_{iBe}},$$

where ℓ_{iAe} and ℓ_{iBe} are the reported likelihoods for Bags A and B. I then re-estimate Equation (19) in L/P using $p_e - \tilde{p}_{ie}$ as the dependent variable. Comparing the resulting coefficients with those for reported posterior bias shows how much of the likelihood-level effect is reproduced by participants' subjective relative evidence.

Column (3) shows that likelihood-implied posterior bias responds in the same direction as reported posterior bias, although the magnitudes are smaller. Implied bias is 0.74 percentage points lower in high- than in low-likelihood environments ($p < 0.05$), compared with 5.17 percentage points for reported beliefs; the medium–low difference is small and statistically insignificant. Whereas the main experiment documents close alignment between reported and likelihood-implied posteriors, this matched comparison reveals that reported likelihoods account for only a limited part of the corresponding change in reported beliefs. Appendix Figure C.3 illustrates this by plotting reported and likelihood-implied posteriors for each objective ratio.

Overall, these findings reject ratio invariance: absolute likelihood levels systematically affect reported posterior beliefs even when objective relative evidence is unchanged, with greater under-inference at low likelihood levels and even reversals to over-inference at high likelihood levels in two conditions. This is directly contrary to Bayesian updating and ratio-compression behavioral models, which predict no such differences or systematic dependence under fixed

objective ratios. The next comparison asks whether holding both objective likelihoods fixed is sufficient to hold subjective evaluations fixed, or whether sample composition also matters.

5.5 Sample Composition and Sample Invariance

The preceding results show that the objective likelihood ratio does not fully characterize subjective evidence. I next examine whether the full objective likelihood pair is sufficient, as formalized in Proposition 2. Panel B of Figure 9 compares a pure two-draw sample with a mixed two-draw sample ($G = 2$ and $G = 1$), both yielding objective likelihoods $(0.16, 0.09)$ and the same Bayesian posterior. This comparison isolates the role of sample composition beyond its implications for the objective likelihood pair.

Restricting the analysis to these environments, I regress the likelihood and posterior bias measures from Equations (18) and (19) on a mixed-sample indicator, retaining bag fixed effects in the likelihood specification and clustering standard errors by participant. Column (1) of Table 8 shows that likelihood bias is 3.80 percentage points greater for the mixed sample (SE = 1.89), rejecting likelihood-pair sufficiency. Posterior under-inference is also greater for the mixed sample in L/P by 3.52 percentage points (SE = 1.63), while the difference in P is smaller and statistically insignificant (1.24; SE = 1.44). Column (4) shows that the likelihood-implied posterior estimate points in the opposite direction and is statistically insignificant (-2.05; SE = 1.30). The clearest finding is therefore that sample composition affects subjective likelihood judgments even when both objective likelihoods are held fixed; its implications for posterior beliefs are less consistent.

Table 8. Reporting Bias by Sample Composition

	(1) Likelihood	(2) Posterior (P)	(3) Posterior (L/P)	(4) Implied Posterior (L/P)
Mixed sample	3.80**	1.24	3.52**	-2.05
	(1.89)	(1.44)	(1.63)	(1.30)
N	372	214	186	186
R^2	0.015	0.002	0.022	0.013

Notes: The sample is restricted to the pure- and mixed-sample benchmark environments in Panel B. The coefficient reports the difference between the mixed and pure environments, with the pure environment as the omitted category. Column (1) uses likelihood-reporting bias, defined as the reported likelihood minus the true likelihood, and pools observations from Bags A and B with bag fixed effects. Columns (2) and (3) defines posterior bias as the Bayesian posterior minus the reported posterior for P and L/P respectively, while Column (4) defines posterior bias as the Bayesian posterior minus the implied posteriors from reported likelihoods in L/P. Standard errors clustered at the participant level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.6 Discussion

The second experiment shows that absolute likelihood levels systematically shape belief updating even when objective relative support is unchanged. Within matched likelihood-ratio conditions, reported beliefs exhibit greater under-inference at low likelihood levels; at higher levels, this bias diminishes and sometimes reverses to over-inference. This violates a core invariance property shared by Bayesian updating and behavioral models that apply a stable transformation to the objective likelihood ratio. Explanations based only on sample proportion and sample size, features emphasized in the strength–weight distinction examined by Griffin and Tversky (1992), also cannot account for these differences because both remain fixed within each comparison.

Observing likelihood judgments further reveals how departures from ratio invariance already emerge during evidence evaluation. Estimating compression or amplification of the objective likelihood ratio from posterior reports to explain belief biases – which is commonly done in much existing work on belief updating – cannot distinguish distorted relative evidence from errors in combining that evidence with priors. By observing reported likelihoods directly, this experiment shows that lower likelihood levels generate greater likelihood overestimation and, for two-draw samples, greater compression of median subjective likelihood ratios toward one, which systematically affect posterior beliefs. A stable compression of the objective ratio alone cannot explain these results, suggesting that parameters commonly interpreted as underreaction or overreaction in this literature may additionally reflect distortions in the likelihood inputs.

6 Conclusion

This paper identifies evidence evaluation as an important source of systematic errors in belief updating. By separating likelihood judgments from posterior beliefs, I show that substantial distortions arise during evaluating evidence before it is combined with priors. These distortions then carry through with little additional error and account for much of the observed posterior under-inference: reported posterior beliefs closely track the posteriors implied by participants' own likelihood reports, while independently elicited likelihood judgments reproduce much of the under-inference observed in the baseline treatment. Intervening in the likelihood stage through reporting or providing likelihoods additionally improves posterior formation, closing approximately 18–20% of the gap to the Bayesian benchmark.

The second experiment establishes absolute likelihood levels as a systematic determinant of belief formation beyond objective relative evidence. Holding the objective likelihood ratio fixed,

lower likelihood levels generate greater likelihood overestimation and, for two-draw samples, greater compression of median subjective likelihood ratios toward one. Reported posteriors also exhibit greater under-inference at lower levels, with errors reversing to over-inference at higher levels in some conditions. These findings reject the ratio-invariance restriction shared by Bayesian updating and behavioral models based on a stable transformation of the objective likelihood ratio.

These findings show why the intermediate evaluation of evidence matters for both future empirical and theoretical work on belief formation. For empirical work, the results change what can be inferred from observed posterior errors. Parameters commonly interpreted as underreaction or overreaction may combine distortions in likelihood inputs with departures from Bayesian posterior formation. Posterior reports alone cannot determine their respective contributions. Moreover, estimates that summarize responsiveness to the objective likelihood ratio may not generalize across environments with different absolute likelihood levels. Separately eliciting likelihood judgments makes these inputs observable and permits direct tests of whether posterior errors are consistent with the evidence participants report evaluating.

For theoretical work, the joint observations impose restrictions that posterior reports alone leave unresolved. A model must explain both how subjective evidence varies with absolute likelihood scale and how that evidence relates to final beliefs. It must also accommodate the close correspondence between reported and likelihood-implied posteriors in the first experiment alongside their different responses to likelihood levels in the second. Distorted sampling beliefs, representativeness, attention, and cognitive imprecision offer candidate mechanisms whose implications can be assessed against both sets of judgments.

These distinctions also matter for the design and evaluation of interventions (e.g., Kling et al., 2012; Ambuehl et al., 2022; Bernheim et al., 2026). Assistance may improve the construction of likelihood judgments, their use in posterior formation, or both. The treatment results show that making likelihoods explicit improves updating, while the remaining under-inference when correct likelihoods are supplied underscores the importance of how those inputs are used. Extending this approach to expectations formation, financial decisions, consumer learning, and hiring would help identify which stage requires assistance and whether improvements in belief formation translate into better decisions.

References

Afrouzi, Hassan, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar, "Overreaction in expectations: Evidence and theory," *The Quarterly Journal of Economics*, 2023, *138* (3), 1713–1764.

Aina, Chiara, Andrea Amelio, and Katharina Brütt, "Contingent belief updating," Technical Report, ECONtribute Discussion Paper 2023.

Ambuehl, Sandro, B Douglas Bernheim, and Annamaria Lusardi, "Evaluating deliberative competence: A simple method with an application to financial choice," *American Economic Review*, 2022, *112* (11), 3584–3626.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from weak signals and underinference from strong signals," *The Quarterly Journal of Economics*, 2025, *140* (1), 335–401.

Ba, Cuimin, J Aislinn Bohren, and Alex Imas, "Over-and underreaction to information," *Available at SSRN 4274617*, 2024.

Bar-Hillel, Maya, "On the subjective probability of compound events," *Organizational behavior and human performance*, 1973, *9* (3), 396–406.

_ , "The base-rate fallacy in probability judgments," *Acta Psychologica*, 1980, *44* (3), 211–233.

Bastianello, Francesca and Alex Imas, "Biases in belief updating within and across domains," Technical Report, National Bureau of Economic Research 2026.

Benjamin, Dan, Aaron Bodoh-Creed, and Matthew Rabin, "Base-rate neglect: Foundations and implications," *Base-Rate Neglect: Foundations and Implications*, 2019, *1*, 62.

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

_ **, Don A Moore, and Matthew Rabin**, "Biased beliefs about random samples: Evidence from two integrated experiments," Technical Report, National Bureau of Economic Research 2017.

_ **, Matthew Rabin, and Collin Raymond**, "A model of nonbelief in the law of large numbers," *Journal of the European Economic Association*, 2016, *14* (2), 515–544.

Bernheim, B Douglas, Aldo Lucia, Kirby Nielsen, and Charles D Sprenger, "When Are Decisions Improvable? An Evaluation of Diagnostic Methods," Technical Report, National Bureau of Economic Research 2026.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg, "The dynamics of discrimination: Theory and evidence," *American economic review*, 2019, *109* (10), 3395–3436.

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

_ **, Nicola Gennaioli, and Andrei Shleifer**, "Overreaction and diagnostic expectations in macroeconomics," *Journal of Economic Perspectives*, 2022, *36* (3), 223–244.

_ **,** _ **, Yueran Ma, and Andrei Shleifer**, "Overreaction in macroeconomic expectations," *American Economic Review*, 2020, *110* (9), 2748–2782.

Camerer, Colin F, "Do biases in probability judgment matter in markets? Experimental evidence," *The American Economic Review*, 1987, *77* (5), 981–997.

Coibion, Olivier and Yuriy Gorodnichenko, "Information rigidity and the expectations formation process: A simple framework and new facts," *American Economic Review*, 2015, *105* (8), 2644–2678.

Danz, David, Lise Vesterlund, and Alistair J Wilson, "Belief elicitation and behavioral incentive compatibility," *American Economic Review*, 2022, *112* (9), 2851–2883.

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

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

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel, "Mental models and learning: The case of base-rate neglect," *American Economic Review*, 2024, *114* (3), 752–782.

Fan, Tony Q, Yucheng Liang, and Cameron Peng, *The inference-forecast gap in belief updating*, LSE Financial Markets Group, 2024.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind," *The Quarterly Journal of Economics*, 2010, *125* (4), 1399–1433.

Giglio, Stefano and Bryan Kelly, "Excess volatility: Beyond discount rates," *The Quarterly Journal of Economics*, 2018, *133* (1), 71–127.

Gonçalves, Duarte, Jonathan Libgober, and Jack Willis, "Retractions: Updating from complex information," *Review of Economic Studies*, 2026, *93* (1), 476–516.

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

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

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

Holt, Charles A and Angela M Smith, "An update on Bayesian updating," *Journal of Economic Behavior & Organization*, 2009, *69* (2), 125–134.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive psychology*, 1972, *3* (3), 430–454.

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

Khaw, Mel Win, Luminita Stevens, and Michael Woodford, "Discrete adjustment to a changing environment: Experimental evidence," *Journal of Monetary Economics*, 2017, *91*, 88–103.

Kling, Jeffrey R, Sendhil Mullainathan, Eldar Shafir, Lee C Vermeulen, and Marian V Wrobel, "Comparison friction: Experimental evidence from Medicare drug plans," *The quarterly journal of economics*, 2012, *127* (1), 199–235.

Mullainathan, Sendhil and Ziad Obermeyer, "Diagnosing physician error: A machine learning approach to low-value health care," *The Quarterly Journal of Economics*, 2022, *137* (2), 679–727.

Prat-Carrabin, Arthur and Michael Woodford, "Imprecise probabilistic inference from sequential data.," *Psychological Review*, 2024.

Rabin, Matthew, "Inference by believers in the law of small numbers," *The Quarterly Journal of Economics*, 2002, *117* (3), 775–816.

_ , "Incorporating limited rationality into economics," *Journal of Economic Literature*, 2013, *51* (2), 528–543.

_ **and Dimitri Vayanos**, "The gambler's and hot-hand fallacies: Theory and applications," *The Review of Economic Studies*, 2010, *77* (2), 730–778.

Tversky, Amos and Daniel Kahneman, "Belief in the law of small numbers.," *Psychological bulletin*, 1971, *76* (2), 105.

_ **and** _ , "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, 1974, *185* (4157), 1124–1131.

_ **and** _ , "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.," *Psychological review*, 1983, *90* (4), 293.

A Experiment 1: Supplementary Results

A.1 Sample Characteristics

Table A.1. Summary Statistics

	Full Sample	P	L/P	L*/P	L
Number of participants	800	207	187	203	203
Demographic information					
Age	43.7	44.5	43.9	44.1	42.2
Male	54.1	54.6	55.1	55.7	51.2
White	75.2	72.5	76.5	78.3	73.9
English first language	96.0	96.6	95.7	95.1	96.6
Current student	7.4	3.9	10.2	7.4	8.4
Full-time employed	51.0	51.2	52.4	50.2	50.2
Born in U.S.	93.5	92.8	96.8	91.6	93.1
Prior Prolific approvals	4693.9	4907.7	4378.7	4571.9	4888.1
Time taken					
Total time taken (minutes)	16.4	13.8	19.8	17.0	15.3
Each round (seconds)	29.9	20.9	42.3	25.0	32.5
Comprehension checks					
Drawing with replacement correct	92.4	92.3	92.5	93.6	91.1
Drawing order correct	85.5	83.6	85.6	86.7	86.2
Likelihood task correct	70.7	–	71.1	62.1	78.8
Posterior task correct	69.0	84.1	63.1	59.1	–

Notes: The table reports group means by treatment condition. Percentages are reported for binary variables. Treatment labels indicate whether likelihoods and/or posterior beliefs were reported; an asterisk denotes that likelihoods were provided rather than reported. Comprehension-check variables are blank in treatments where the corresponding belief object was not elicited (e.g., posterior-belief checks are not asked in the likelihood-only treatment, and likelihood checks are not asked in the posterior-only treatment).

A.2 Likelihood Judgments by Signal Environment

Table A.2. Reported Likelihoods by Signal Environment and Bag

	Bag 1				Bag 2			
	Objective	L/P	L	L/P − L	Objective	L/P	L	L/P − L
Panel A: $M = 1$								
$G = 0$	29.94	31.71	30.55	-1.15	70.06	67.56	68.74	1.18
				(1.00)				(1.25)
$G = 1$	70.14	70.12	69.12	-1.00	29.86	30.09	30.12	0.03
				(1.11)				(1.00)
Panel B: $M = 2$								
$G = 0$	9.16	25.20	22.60	-2.60	49.18	63.86	61.01	-2.85
				(1.66)				(1.78)
$G = 1$	20.81	43.85	39.99	-3.86*	20.81	43.14	38.37	-4.77**
				(2.02)				(1.94)
$G = 2$	49.09	63.47	62.98	-0.49	9.19	23.58	22.17	-1.40
				(1.76)				(1.36)
Panel C: $M = 3$								
$G = 0$	2.91	21.70	17.83	-3.87*	34.31	58.85	54.39	-4.46
				(2.19)				(2.76)
$G = 1$	6.29	26.78	24.57	-2.22	14.52	53.93	47.08	-6.85**
				(2.14)				(3.27)
$G = 2$	14.53	52.98	46.32	-6.65**	6.35	27.76	22.16	-5.60***
				(3.10)				(2.06)
$G = 3$	34.76	61.14	56.80	-4.35*	2.84	19.84	17.85	-1.99
				(2.52)				(1.61)

Notes: Each row corresponds to a signal environment defined by the number of draws (M) and the number of green draws (G). Columns report the objective benchmark and mean stated likelihoods separately by treatment and bag. L/P denotes the treatment in which participants reported both likelihoods and posteriors; L denotes the treatment in which participants reported likelihoods only. The fourth and eighth columns report the difference in mean reported likelihoods, L/P minus L, with cluster-robust standard errors in parentheses.

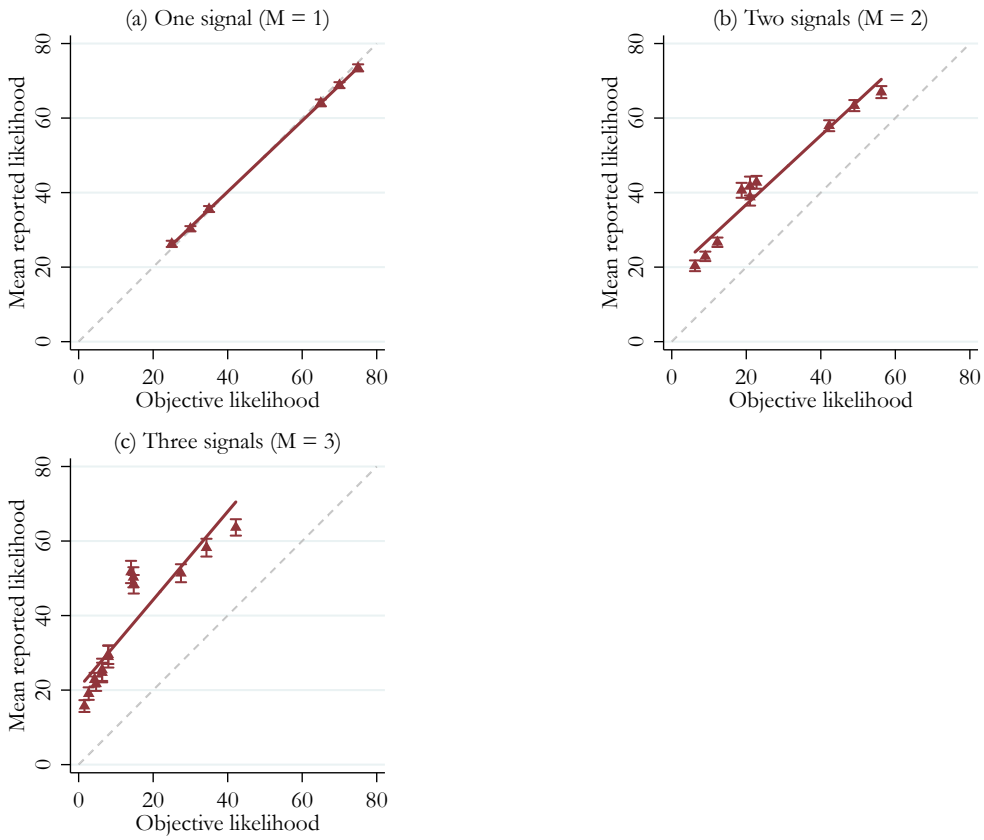


Figure A.1. Reported versus Objective Likelihoods by Number of Signals

Notes: Panels (a)–(c) correspond to $M = 1, 2, 3$ signals. Each panel plots mean reported likelihoods against the objective benchmark, pooling across bags and treatments. Points represent means with 95% confidence intervals. The dashed line denotes the 45-degree identity line; solid lines are linear fits.

A.2.1 Likelihood Alignment by Number of Signals

To examine how likelihood distortions vary with sample size, I estimate a specification allowing both the intercept and slope to differ across $M = 1, 2, 3$, separately by treatment and bag:

$$l_{ibt} = \alpha_{gb} + \sum_{m=2,3} \eta_{gbm} \mathbf{1}\{M = m\} + \beta_{gb} L_{ibt} + \sum_{m=2,3} \gamma_{gbm} L_{ibt} \mathbf{1}\{M = m\} + \varepsilon_{ibt}, \tag{20}$$

where l_{ibt} and L_{ibt} denote the reported and objective likelihoods for participant i, bag b, and round t, and g indexes treatment. The coefficients α_{gb} and β_{gb} give the intercept and slope for single-draw samples. The coefficients η_{gbm} and γ_{gbm} capture the corresponding intercept and slope differences for samples of size m relative to $M = 1$. Because some signal environments occur across treatments, this specification relies on pooled variation and is less directly interpretable than the treatment-specific estimates in the main text.

Table A.3 shows that single-draw likelihood reports have small intercepts and slopes close to one. For two-draw samples, intercepts increase by 14.8–16.6 percentage points across treatments and bags, while slope differences are small and statistically insignificant. For three-draw samples, intercepts increase by 17.2–20.8 percentage points relative to single draws, and slopes increase by 0.19–0.29. The resulting fitted intercepts range from 16.2 to 20.2 for two draws and from 18.3 to 23.4 for three draws.

These estimates show that substantial upward distortions in multiple-draw likelihood judgments persist when both intercepts and slopes vary freely with sample size. For two draws, the departure from single-draw reports is primarily an upward shift; for three draws, it also involves greater responsiveness to objective likelihoods. The distortions therefore cannot be characterized solely by reduced responsiveness to the objective benchmark.

Table A.3. Reported Likelihoods versus Objective Likelihoods by Number of Signals

	(1) L/P, Bag A	(2) L, Bag A	(3) L/P, Bag B	(4) L, Bag B
Intercept: $M = 1$	3.608**	1.480	2.551*	1.113
	(1.465)	(1.152)	(1.344)	(1.371)
Intercept difference: $M = 2$	16.570***	14.820***	16.392***	15.057***
	(1.696)	(1.587)	(1.640)	(1.575)
Intercept difference: $M = 3$	19.561***	18.037***	20.841***	17.213***
	(1.771)	(1.468)	(1.725)	(1.709)
Slope: $M = 1$	0.944***	0.968***	0.928***	0.966***
	(0.026)	(0.024)	(0.027)	(0.028)
Slope difference: $M = 2$	-0.030	-0.004	0.009	-0.045
	(0.039)	(0.040)	(0.037)	(0.036)
Slope difference: $M = 3$	0.294***	0.191***	0.225***	0.191***
	(0.057)	(0.050)	(0.066)	(0.050)
N	2805	3045	2805	3045
R^2	0.50	0.54	0.48	0.56

Notes: Each column reports a regression estimated separately by treatment and bag, allowing both the intercept and slope to vary with sample size. The omitted category is $M = 1$. Intercept and slope differences are relative to this category. Reported and objective likelihoods are expressed in percentage points. Standard errors clustered at the participant level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3 Posterior Beliefs by Signal Environment

Table A.4. Reported Posteriors by Signal Environment

	Bayes	P	L/P	L*/P	P − L/P	L/P − L*/P	P − L*/P
Panel A: $M = 1$							
$G = 0$	30.00	32.90	31.01	32.29	-1.90*	-1.28	-0.61
					(1.06)	(1.03)	(1.03)
$G = 1$	70.11	69.38	70.55	67.86	1.17	2.69***	-1.52*
					(0.83)	(0.81)	(0.81)
Panel B: $M = 2$							
$G = 0$	16.10	29.91	27.00	24.70	-2.90**	2.30	-5.20***
					(1.45)	(1.52)	(1.51)
$G = 1$	50.00	52.04	50.59	48.67	-1.46	1.92*	-3.37***
					(1.29)	(1.12)	(1.01)
$G = 2$	83.84	72.57	73.64	72.86	1.07	0.79	0.28
					(1.06)	(1.31)	(1.22)
Panel C: $M = 3$							
$G = 0$	8.17	25.63	23.76	18.42	-1.87	5.34***	-7.20***
					(2.05)	(1.89)	(1.70)
$G = 1$	30.17	35.02	31.83	32.74	-3.19**	-0.91	-2.28
					(1.58)	(1.66)	(1.53)
$G = 2$	70.15	67.91	67.06	65.22	-0.85	1.84	-2.69**
					(1.45)	(1.52)	(1.28)
$G = 3$	91.74	76.35	77.44	77.62	1.09	-0.17	1.27
					(1.50)	(1.88)	(1.74)

Notes: Each row corresponds to a signal environment defined by the number of draws (M) and the number of green draws (G). Columns report the Bayesian benchmark and mean stated posteriors by treatment. P denotes the treatment in which participants reported posteriors only; L/P denotes the treatment in which participants reported both likelihoods and posteriors; L*/P denotes the treatment in which likelihoods were provided before participants reported posteriors. The final three columns report pairwise differences in mean posterior reports, with cluster-robust standard errors in parentheses.

These estimates show that substantial upward distortions in multiple-draw likelihood judgments persist when both intercepts and slopes vary freely with sample size. For two draws, the departure from single-draw reports is primarily an upward shift; for three draws, it also involves greater responsiveness to objective likelihoods. The distortions therefore cannot be characterized solely by reduced responsiveness to the objective benchmark.

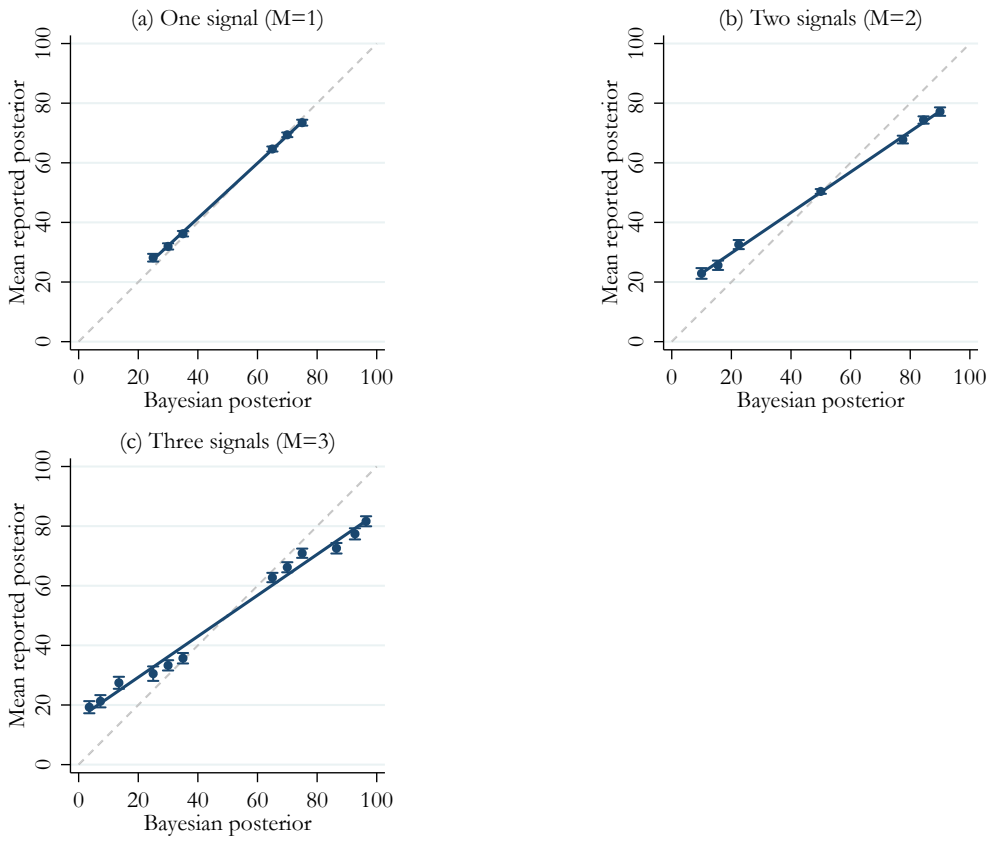


Figure A.2. Reported versus Bayesian Posteriors by Number of Signals

Notes: Panels (a)–(c) correspond to $M = 1, 2, 3$ signals. Each panel plots mean reported posteriors against the corresponding Bayesian posterior, pooling across treatments. Points represent means with 95% confidence intervals. The dashed line denotes the 45-degree identity line; solid lines are linear fits.

A.3.1 Posterior Alignment by Number of Signals

I estimate a joint specification allowing alignment with the Bayesian benchmark to vary with the number of signals M:

$$\pi_{it} - 50 = \alpha_g + \beta_g(P_{it} - 50) + \sum_{m=2,3} \gamma_{gm}(P_{it} - 50)\mathbf{1}\{M = m\} + \varepsilon_{it}. \tag{21}$$

where the coefficient on $(P_{it} - 50)$ gives the slope for single-signal observations ($M = 1$), and the interaction terms capture slope differences for $M = 2$ and $M = 3$ relative to $M = 1$. Because some signal environments occur across treatments, this specification relies on pooled variation and is therefore less directly interpretable than the treatment-specific estimates in the main text.

Table A.5 reports the estimates. Slopes are close to one for single-signal observations across treatments and differ modestly across larger signal environments, without altering the main conclusion that deviations from the Bayesian benchmark arise primarily when multiple signals must be aggregated.

Table A.5. Reported Posteriors versus Bayesian Posteriors by Number of Signals

	(1) P	(2) L/P	(3) L*/P
Base slope ($M = 1$)	0.908***	0.977***	0.886***
	(0.028)	(0.025)	(0.025)
Slope difference: $M = 2$	-0.276***	-0.285***	-0.174***
	(0.027)	(0.028)	(0.029)
Slope difference: $M = 3$	-0.263***	-0.281***	-0.160***
	(0.030)	(0.031)	(0.029)
Constant	0.997**	0.660	0.176
	(0.413)	(0.418)	(0.354)
N	3105	2805	3045
R^2	0.68	0.71	0.72

Notes: This table reports estimates from Equation (21). The omitted category is $M = 1$, so the coefficient on Bayesian posterior (centered) gives the slope for single-signal observations, and the interaction terms report slope differences for $M = 2$ and $M = 3$ relative to $M = 1$. The dependent variable and Bayesian posterior are centered at 50. Standard errors are clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.4 Reported and Likelihood-Implied Posteriors

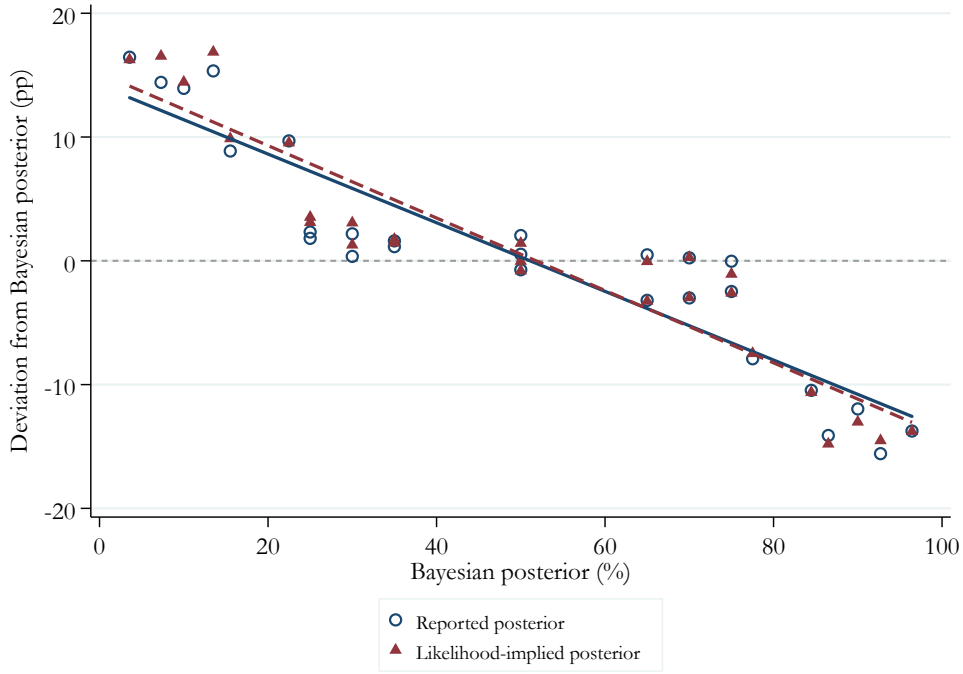


Figure A.3. Reported and Likelihood-Implied Posterior Errors in the L/P Treatment

Notes: Each pair of points corresponds to a signal environment (M, δ, G). The horizontal axis plots the objective Bayesian posterior. The vertical axis reports mean deviations from this benchmark, in percentage points, for reported posteriors (circles) and posteriors implied by participants' own likelihood judgments (triangles). Implied posteriors are calculated for each observation before averaging within environments. The solid and dashed fitted lines correspond to reported and likelihood-implied errors, respectively; the horizontal line marks zero error. The difference between the two series measures the mean deviation of reported posteriors from the conditional Bayesian benchmark.

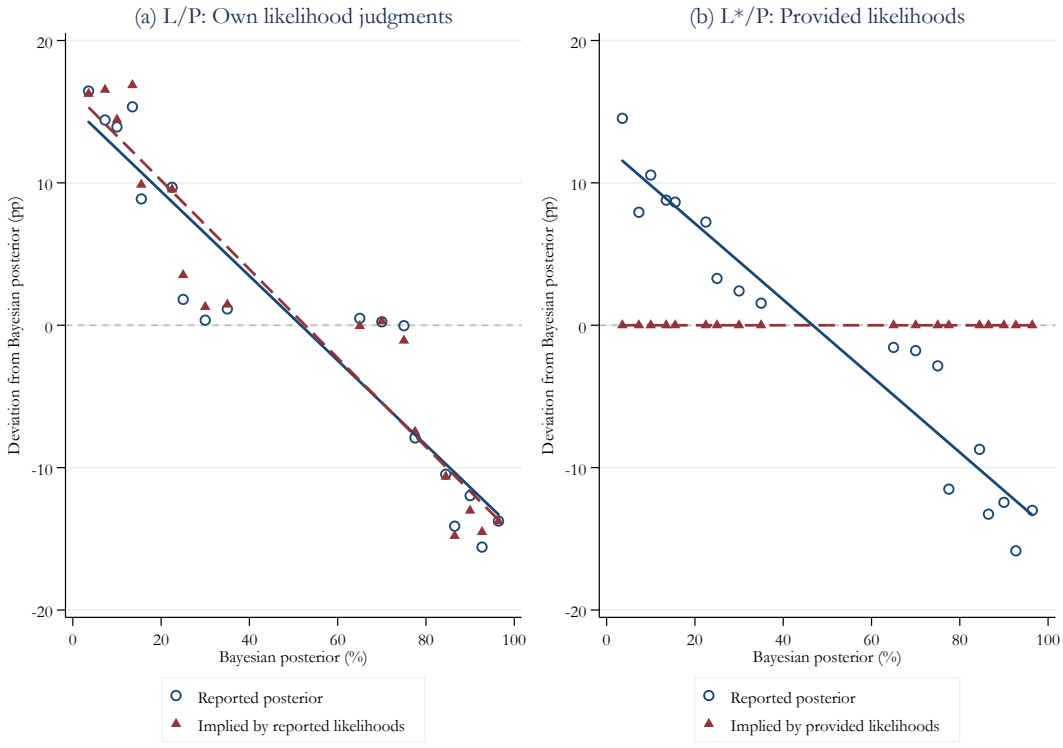


Figure A.4. Reported and Likelihood-Implied Posterior Errors in L/P versus L*/P Treatments

Notes: Each point represents a signal environment (M, δ, G). The horizontal axis shows the objective Bayesian posterior; the vertical axis shows mean reported posterior minus this benchmark, in percentage points. Circles show reported-posterior deviations. Triangles show deviations of the posterior implied by participants' own likelihood reports in L/P and, in L*/P, the posterior implied by the likelihoods provided to participants. The dashed horizontal line marks zero deviation. The fitted lines summarize reported-posterior deviations by treatment. The distance between a treatment's reported-posterior deviation and its implied-posterior deviation measures the mean difference between reported and implied posteriors.

A.5 Provided Likelihoods: Implementation and Robustness

The experimental task displayed ordered sequences, but the likelihood values supplied in L*/P were inadvertently calculated for unordered sample compositions. For a sequence of M draws containing G green draws, the supplied values exceeded the correct sequence likelihoods by the common factor $\binom{M}{G}$. For example, the supplied likelihood for a particular sequence containing two green draws and one orange draw was three times its objective likelihood. Although this factor cancels from the likelihood ratio and leaves the implied Bayesian posterior unchanged, the supplied values did not correctly describe the probability of the exact displayed sequence. The discrepancy affects only mixed samples with $M > 1$, which account for 27.6% of observations in L*/P. Single-draw and pure multiple-draw samples have $\binom{M}{G} = 1$ and are unaffected. The main analysis therefore excludes the affected observations.

To examine behavior across these environments, I re-estimate Equation (9) separately by treatment for unaffected environments where the correct likelihood values were provided, comprising single draws and pure multiple-draw sequences, and affected environments, comprising mixed multiple-draw sequences where incorrect likelihoods were supplied. I apply the same environment restrictions to P and L/P, allowing treatment comparisons within each group.

Columns (1)–(3) of Table A.6 report the estimates for unaffected environments. The slope is 0.746 in L*/P, compared with 0.735 in L/P and 0.677 in P. Thus, the greater responsiveness associated with eliciting or providing likelihoods is also present when all three treatments are compared using the environments in which the supplied likelihoods were correct. In contrast, Columns (4)–(6) report estimates for affected environments. Here, the slope in L*/P is 0.807, nearly identical to P (0.807; $p = 0.99$) and lower than L/P (0.880), although the latter difference is not statistically significant ($p = 0.21$). These estimates do not show an improvement in responsiveness from providing likelihoods within the affected environments.

These patterns indicate that the implementation issue is concentrated in mixed samples in L*/P, where ordered and unordered likelihoods differ, because of which I exclude these from the main analysis. However, because participants completed multiple rounds, exposure to the incorrectly displayed likelihoods in some rounds could have influenced responses in others. Excluding affected observations therefore removes reports made directly after incorrectly supplied likelihoods, but cannot exclude any possible effects of that exposure on responses in other rounds for this treatment group.

Table A.6. Posterior Updating in Unaffected and Affected Environments

	Unaffected environments			Affected environments		
	(1)	(2)	(3)	(4)	(5)	(6)
	P	L/P	L*/P	P	L/P	L*/P
Bayesian posterior	0.677***	0.735***	0.746***	0.807***	0.880***	0.807***
	(0.015)	(0.019)	(0.019)	(0.037)	(0.042)	(0.039)
Constant	0.989**	0.667	-0.656	1.489***	-0.063	-1.250**
	(0.398)	(0.432)	(0.399)	(0.470)	(0.653)	(0.487)
N	2268	2002	2203	837	803	842
R^2	0.71	0.75	0.74	0.46	0.48	0.52

Notes: Each column reports a regression of reported posterior beliefs on the corresponding Bayesian posterior, estimated separately by treatment. Unaffected environments comprise single draws and pure multiple-draw sequences, for which the likelihoods supplied in L*/P were correct. Affected environments comprise mixed multiple-draw sequences, for which the supplied likelihoods exceeded the correct sequence probabilities by a common combinatorial factor. The same environment restrictions are applied to P and L/P. Both reported and Bayesian posteriors are centered at 50. Standard errors clustered at the participant level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.6 Pre-Registered Sample Filters

As specified in the pre-analysis plan, I examine robustness to the following sample filters, applied independently to the full sample:

1. Participants must complete all rounds and pass a majority of the comprehension checks in their assigned treatment. This excludes 122 of 800 participants.

2. Participants with implausibly fast completion times (less than 7 minutes) are excluded. This excludes 74 participants.

3. Participants must exhibit positive slopes in individual-level regressions of reported posteriors on the corresponding Bayesian values, ensuring meaningful updating behavior. This excludes 6 participants from L, L/P, and L*/P; L is not applicable because posteriors were not elicited.

Applying these filters cumulatively yields a restricted sample of 611 participants (an exclusion of 23.6%), below the 30% cutoff specified in the pre-analysis plan. Tables A.7 and A.8 report the main likelihood and posterior specifications using the restricted sample, where the former is restricted to $M > 1$ for comparability with the main analysis. Tables A.9 and A.10 report the pooled specifications by number of signals.

Table A.7. Reported Likelihoods versus Objective Likelihoods by Treatment and Bag (Restricted Sample)

	(1) L/P Treatment	(2) L Treatment	(3) Pooled
True likelihood	0.917***	0.946***	0.931***
	(0.033)	(0.023)	(0.019)
Constant	23.056***	18.229***	20.395***
	(1.202)	(1.056)	(0.808)
N	2716	3488	6204
R^2	0.33	0.41	0.37

Notes: Columns report linear regressions of reported likelihoods on the corresponding objective likelihood, estimated separately by treatment and bag, using the restricted sample that applies the cumulative pre-registered filters. The dependent variable is the reported likelihood, and the independent variable is the objective likelihood. In L/P, participants reported likelihoods and posteriors; in L, participants reported likelihoods only. Unit of observation is participant–round–bag. Standard errors clustered at the participant level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8. Reported Posteriors versus Bayesian Posteriors by Treatment (Restricted Sample)

	(1) P	(2) L/P	(3) L*/P	(4) Pooled
Bayesian posterior (centered)	0.697***	0.762***	0.777***	0.744***
	(0.018)	(0.022)	(0.021)	(0.012)
Constant	1.158***	0.538	-0.647	0.367
	(0.393)	(0.449)	(0.449)	(0.249)
N	2295	2100	2115	6510
R^2	0.67	0.71	0.75	0.71

Notes: Each column reports a linear regression of reported posterior beliefs on the corresponding Bayesian posterior, estimated separately by treatment group, using the restricted sample that applies the cumulative pre-registered filters. The dependent variable is the participant's reported posterior centered at 50, and the independent variable is the Bayesian posterior centered at 50. The slope coefficient measures responsiveness to the Bayesian benchmark; a value of 1 corresponds to perfect Bayesian updating. Unit of observation is participant–round. Standard errors, clustered at the participant level, are reported in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9. Reported Likelihoods versus Objective Likelihoods by Number of Signals (Restricted Sample)

	(1) L/P, Bag A	(2) L, Bag A	(3) L/P, Bag B	(4) L, Bag B
Base slope ($M = 1$)	0.956***	0.971***		
	(0.028)	(0.026)		
Slope difference: $M = 2$	-0.049	0.010		
	(0.046)	(0.037)		
Slope difference: $M = 3$	0.260***	0.193***		
	(0.069)	(0.051)		
Base slope ($M = 1$)			0.933***	0.963***
			(0.030)	(0.031)
Slope difference: $M = 2$			0.020	-0.037
			(0.042)	(0.039)
Slope difference: $M = 3$			0.234***	0.178***
			(0.076)	(0.050)
Constant	2.480*	1.218	2.829*	1.263
	(1.437)	(1.174)	(1.614)	(1.493)
N	2100	2655	2100	2655
R^2	0.49	0.57	0.49	0.57

Notes: This table reports estimates from Equation (20), estimated separately by treatment and bag, using the restricted sample that applies the cumulative pre-registered filters. The omitted category is $M = 1$, so the coefficient on the objective likelihood gives the slope for single-signal observations, and the interaction terms report slope differences for $M = 2$ and $M = 3$ relative to $M = 1$. Standard errors clustered at the participant level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.10. Reported Posteriors versus Bayesian Posteriors by Number of Signals (Restricted Sample)

	(1) P	(2) L/P	(3) L*/P
Base slope ($M = 1$)	0.923***	0.987***	0.898***
	(0.031)	(0.027)	(0.023)
Slope difference: $M = 2$	-0.278***	-0.274***	-0.177***
	(0.031)	(0.030)	(0.032)
Slope difference: $M = 3$	-0.277***	-0.288***	-0.126***
	(0.035)	(0.035)	(0.031)
Constant	0.850*	0.349	0.667
	(0.443)	(0.422)	(0.410)
N	2295	2100	2115
R^2	0.69	0.72	0.75

Notes: This table reports estimates from Equation (21), using the restricted sample that applies the cumulative pre-registered filters.. The omitted category is $M = 1$, so the coefficient on Bayesian posterior (centered) gives the slope for single-signal observations, and the interaction terms report slope differences for $M = 2$ and $M = 3$ relative to $M = 1$. The dependent variable and Bayesian posterior are centered at 50. Standard errors are clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Experiment 1: Individual Differences

B.1 Individual Heterogeneity

The results in the paper focus on average behavior, which may mask substantial heterogeneity in belief updating. To examine this, I estimate participant-level responsiveness to the normative benchmark at each stage of updating.

For likelihood judgments, I estimate Equation (10) separately for each participant, yielding one slope β_i^L and intercept α_i^L per individual. These coefficients summarize how participant i's reported likelihoods respond to the benchmark likelihood and their fitted reporting level. Similarly, for posterior beliefs, I estimate Equation (12) at the participant level, yielding one pair (β_i^P, α_i^P) for each individual, capturing responsiveness to the Bayesian posterior and level bias at neutral evidence.

Figure B.1 plots the distributions of these participant-level coefficients. Panels (a) and (c) show substantial dispersion in slope coefficients for likelihoods and posteriors, indicating considerable heterogeneity in responsiveness to the benchmark. Many participants exhibit slopes below one, consistent with attenuated responses, but with significant dispersion. Panels (b) and (d) show that intercepts also vary across individuals, with greater dispersion in likelihood intercepts and more limited variation in posterior intercepts around neutral evidence.

While responsiveness varies substantially across individuals at each stage, this does not reveal whether heterogeneity is shared across stages. To assess this, I focus on the L/P treatment, where both likelihoods and posterior beliefs are elicited from the same participants, allowing a direct comparison. I focus on slope coefficients, which capture responsiveness to evidence and are directly comparable across stages. For each participant, I estimate a pair of coefficients (β_i^L, β_i^P) using Equations (10) and (12), and examine their relationship across individuals.

Figure B.2 plots posterior responsiveness (β_i^P) against likelihood responsiveness (β_i^L) at the individual level. There is a strong positive relationship: participants who are more responsive to the benchmark at the likelihood stage also tend to be more responsive at the posterior stage. The fitted slope is 0.784 (SE = 0.046), indicating that heterogeneity in updating behavior is largely shared across stages rather than arising independently.

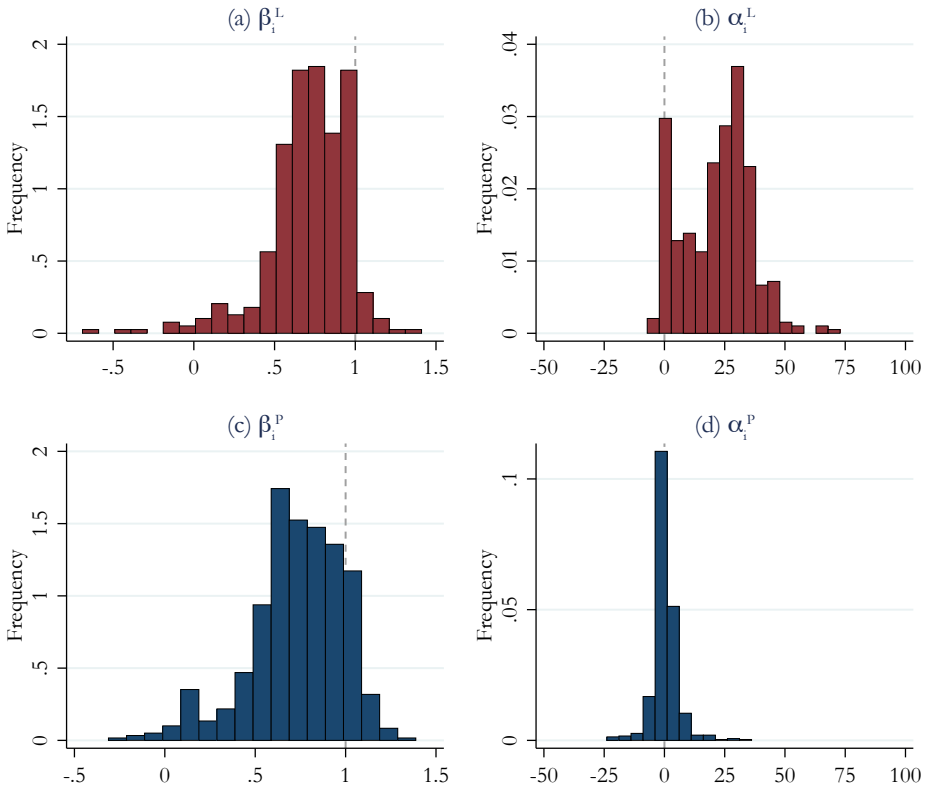


Figure B.1. Individual Responsiveness and Level Differences Across Updating Stages

Notes: Panels (a)–(b) show likelihood coefficients (β_i^L, α_i^L) and Panels (c)–(d) show posterior coefficients (β_i^P, α_i^P), estimated from participant-level regressions of reported beliefs on the corresponding benchmark (Equations (10) and (12)). Dashed lines denote $\beta = 1$ in Panels (a) and (c), and $\alpha = 0$ in Panels (b) and (d).

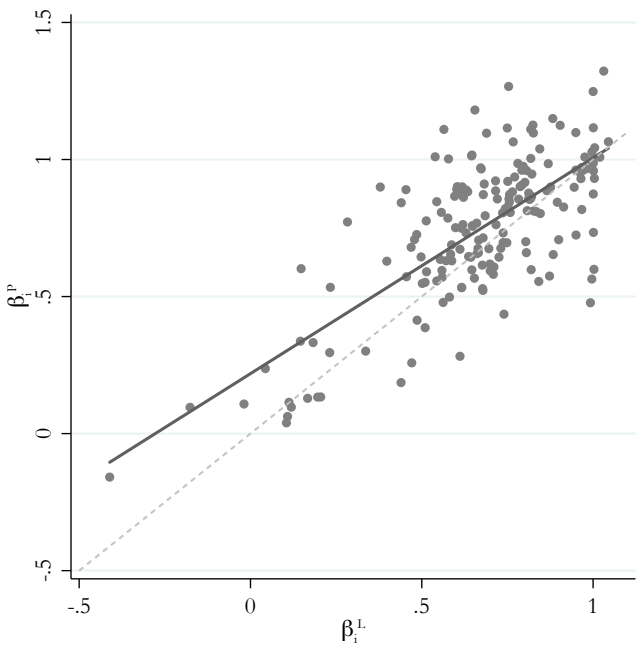


Figure B.2. Relationship Between Likelihood and Posterior Responsiveness Across Participants in the L/P Treatment

Notes: Each point represents a participant in the L/P treatment. The horizontal axis plots the participant-level slope β_i^L from the likelihood regression from (10), and the vertical axis plots β_i^P from the posterior regression from (12). The dashed line denotes the 45-degree identity line, and the solid line shows the linear fit.

B.2 Accuracy and Response Times Across Rounds

I next examine whether behavior changes over time, which would indicate learning or changes in effort. Panel (a) of Figure B.3 shows that mean absolute deviations from the Bayesian posterior (in percentage points) remain stable across rounds, with a regression of deviations on round yielding a small and statistically insignificant coefficient of -0.018 (SE = 0.030). Panel (b) shows that mean response times (in seconds) remain broadly constant over time. Together, these results provide little evidence that learning or changes in effort over the course of the experiment explain the observed results.

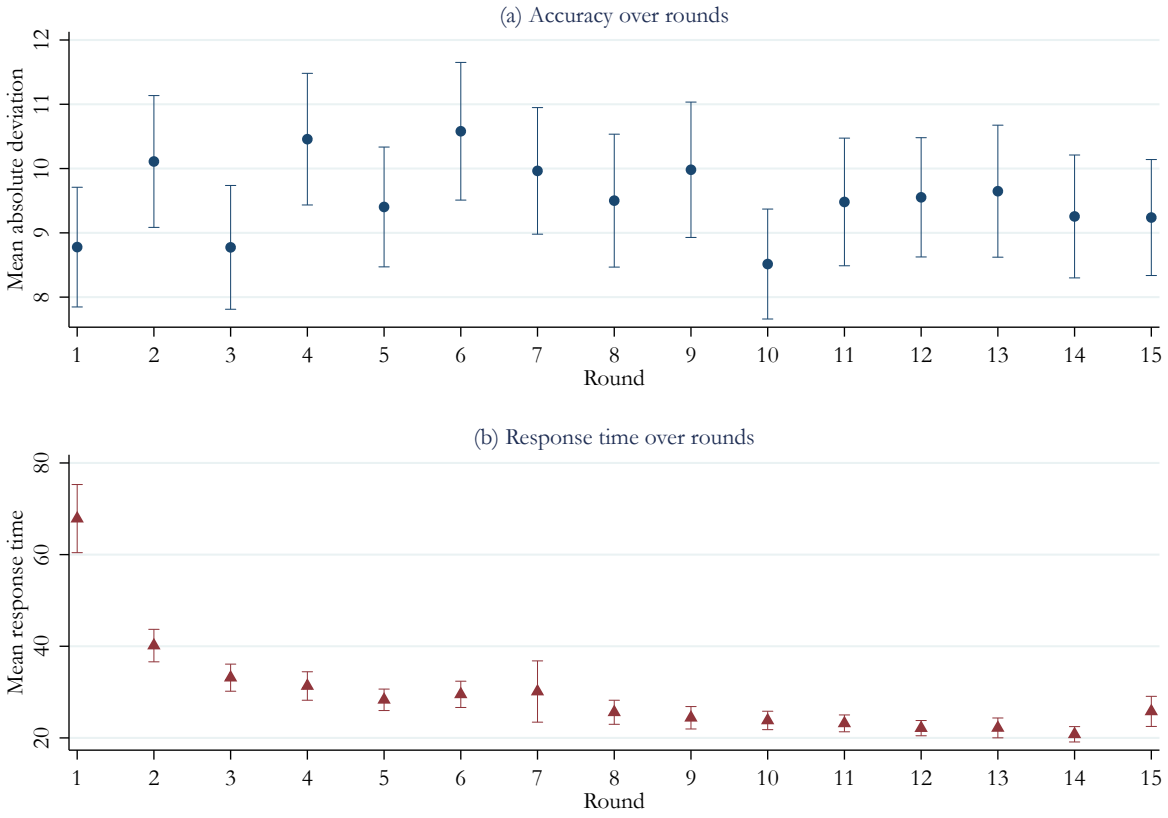


Figure B.3. Accuracy and Response Times over Rounds for Posterior Beliefs

Notes: Panel (a) shows mean absolute deviations from the Bayesian posterior by round, defined as the absolute difference between reported and Bayesian posteriors. Panel (b) shows mean response times (in seconds) by round. The sample includes only treatments in which posterior beliefs are elicited (P, L/P, and L*/P). Error bars denote 95% confidence intervals.

B.3 Response Times Across Signal Environments

Figure B.4 plots mean response times by signal environment, separately for likelihood and posterior tasks. Two patterns stand out. First, response times increase with M, consistent with harder aggregation problems requiring more deliberation. Second, within multi-signal environments ($M > 1$), mixed sequences take significantly longer than pure sequences at both stages (likelihood tasks: 42.5 versus 36.8 seconds; posterior tasks: 33.9 versus 28.5 seconds; $p < 0.001$ in both cases).

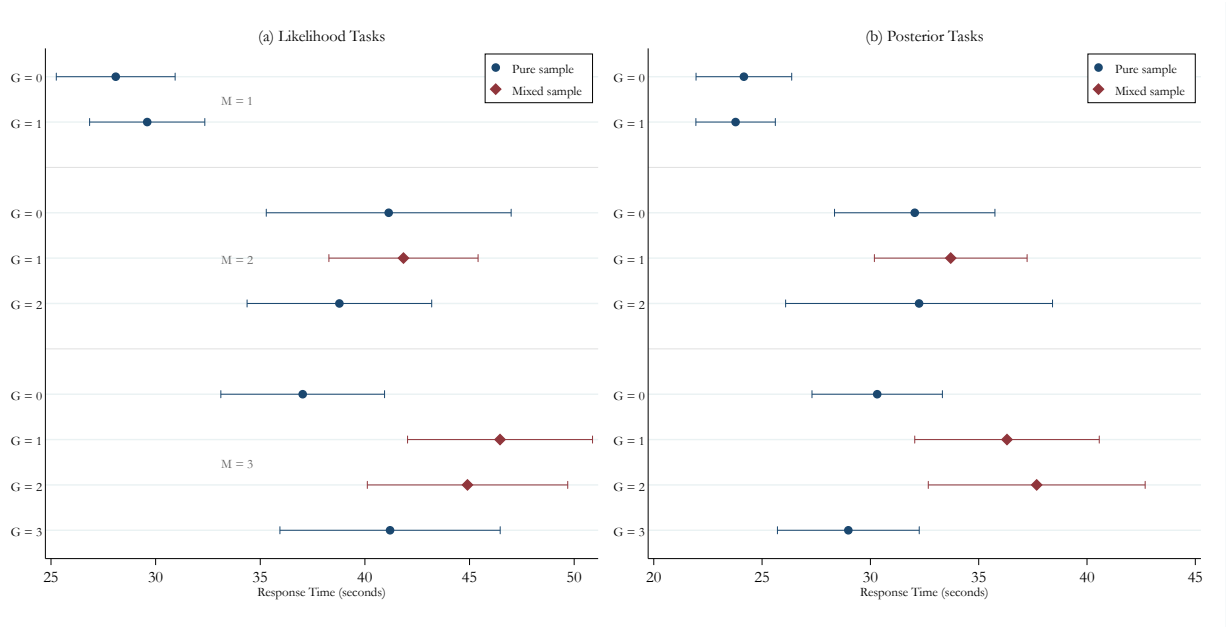


Figure B.4. Response Times by Signal Environment

Notes: Each row corresponds to a signal environment (M, G). Points denote mean response time in seconds, with 95% confidence intervals; maroon points correspond to pure samples ($G = 0$ or $G = M$) and navy points to mixed samples ($0 < G < M$). Panel (a) pools across the L/P and L treatments; Panel (b) pools across the P, L/P, and L*/P treatments.

This pattern might appear to contradict the distortion results in Section 4.5: pure sequences generate larger ratio compression and posterior under-inference, yet participants spend less time on them. One interpretation is that the two environment types involve qualitatively different errors. In pure sequences, one bag's probability of generating the sample is near zero, and participants may quickly substitute a more moderate value—a fast heuristic that produces severe ratio compression without requiring extended deliberation. In mixed sequences, both bags have moderate probabilities yet participants spend more time and still exhibit large absolute distortions, consistent with the concentration of distortions in the lower-probability bag.

B.4 Qualitative Evidence Methodology

To better understand the cognitive processes underlying updating across treatments, participants were asked at the end of the experiment to briefly describe how they formed their probability judgments. The wording of the open-ended question varied slightly by treatment. For example, participants in the posterior-only (P) treatment were asked:

> *"In Which Bag questions, you estimated the chance that each bag produced the sample. How did you make these estimates?"*

Participants in treatments involving likelihood judgments were asked analogous questions referring to their likelihood judgments.

These text responses were first reviewed qualitatively by a large language model (LLM) to identify recurring reasoning patterns within each treatment group. Appendix Table B.1 summarizes the most common themes and illustrative examples. The qualitative patterns reveal a clear progression across treatments: responses in P largely reflect similarity- and proportion-based reasoning, while responses in L/P and L*/P show increasingly structured reasoning involving likelihood comparison and normalization. In contrast, responses in L often demonstrate recognition of the multiplicative structure of sampling with replacement but also frequent difficulty applying it consistently in more complex environments.

Next, to quantify these patterns, the LLM formalized the themes into a structured coding rubric capturing nine distinct reasoning features, including proportion-based heuristics, multiplicative likelihood construction, conditional likelihood comparison, references to priors, normalization procedures, and common computational shortcuts such as averaging or defaulting to 50/50. Each response was then coded using binary, non-exclusive indicators based solely on stated reasoning rather than inferred from numerical answers.

Table B.2 reports the share of responses in each treatment mentioning each feature, providing a quantitative validation of the qualitative patterns. The results reveal systematic differences across treatments that closely track the structure of the task and the demands of likelihood evaluation. In the P treatment, participants rely heavily on proportion-based and visual heuristics (82%), consistent with a large literature showing that individuals evaluate probabilities using representativeness-based heuristics, mapping sample proportions or visual similarity into likelihood judgments rather than applying Bayes' rule (e.g., Kahneman and Tversky, 1972; Tversky and Kahneman, 1974).

In contrast, responses in L/P are more structured: explicit multiplicative likelihood construction rises sharply (28% vs. 6% in P), along with normalization (24%), where participants explicitly compare likelihoods and rescale them to sum to 100. This pattern indicates more systematic processing and an emerging application of Bayes' rule. In L*/P, normalization

Treatment	Common patterns in open-ended responses
P	Proportion-based and visual reasoning (e.g., matching the dominant color); increased confidence with repeated same-color signals; frequent default to 50/50 in mixed or uncertain cases; little explicit probability computation.
L/P	Two-step reasoning: compute likelihoods under each bag (multiplying across draws), then compare and normalize (e.g., "divide by the total"); occasional fallback to simpler heuristics in more complex cases.
L*/P	Mechanical normalization of provided likelihood values; explicit references to Bayes' rule with equal priors; some treatment of provided numbers as cues rather than strict inputs.
L	Recognition of multiplicative structure in sampling with replacement; use of structured calculations; accurate in simple cases but difficulty applying procedures in more complex environments, often leading to averaging or guessing.

Table B.1. Qualitative themes in open-ended responses by treatment.

remains high (30%) but multiplicative construction is rare (3%), consistent with participants applying provided likelihood values rather than generating them independently. In L, participants frequently recognize the multiplicative structure (29.1%) but show weaker normalization and greater inconsistency across environments, suggesting difficulty applying these calculations reliably in more complex settings.

Together, these descriptions show that stated reasoning varies with the judgments participants are asked to make. Likelihood elicitation is associated with more frequent references to multiplicative probability calculations, while posterior elicitation alongside reported or supplied likelihoods is associated with references to normalization. These patterns complement the quantitative results by documenting how participants describe constructing and using likelihood judgments.

	P	**L/P**	**L*/P**	**L**
Proportion-based reasoning	82%	58%	54%	37.9%
Multiplicative likelihoods	6%	28%	3%	29.1%
Conditional likelihood framing	19%	38%	18%	22.7%
Explicit prior mention (50/50)	7%	5%	5%	–
Explicit normalization (ratios)	9%	24%	30%	–
Prior–likelihood integration	3%	12%	8%	–
50/50 fallback	13%	10%	7%	1.5%
Guessing / intuition language	29%	29%	28%	25.6%
Averaging / add-then-divide error	7%	9%	3%	8.9%

Table B.2. Reasoning patterns in open-ended responses by treatment

Notes: Each response is coded for the presence of non-exclusive reasoning features based on participants' written explanations. Entries report the percentage of responses within each treatment that mention the corresponding feature.

C Experiment 2: Design Details and Supplementary Results

C.1 Experimental Environments

Table C.1. Experiment 2: Experimental Environment Parameters

M	G	Level	(δ_A, δ_B)	(L_A, L_B)	Likelihood Ratio	Posterior
Panel A: Likelihood-level manipulation						
1	1	High	$(0.64, 0.36)$	$(0.64, 0.36)$	1.78	64.0%
1	1	Low	$(0.16, 0.09)$	$(0.16, 0.09)$	1.78	64.0%
2	2	High	$(0.80, 0.60)$	$(0.64, 0.36)$	1.78	64.0%
2	2	Low	$(0.40, 0.30)$	$(0.16, 0.09)$	1.78	64.0%
1	1	High	$(0.81, 0.36)$	$(0.81, 0.36)$	2.25	69.2%
1	1	Medium	$(0.36, 0.16)$	$(0.36, 0.16)$	2.25	69.2%
1	1	Low	$(0.09, 0.04)$	$(0.09, 0.04)$	2.25	69.2%
2	2	High	$(0.90, 0.60)$	$(0.81, 0.36)$	2.25	69.2%
2	2	Medium	$(0.60, 0.40)$	$(0.36, 0.16)$	2.25	69.2%
2	2	Low	$(0.30, 0.20)$	$(0.09, 0.04)$	2.25	69.2%
1	1	High	$(0.64, 0.16)$	$(0.64, 0.16)$	4.00	80.0%
1	1	Medium	$(0.36, 0.09)$	$(0.36, 0.09)$	4.00	80.0%
1	1	Low	$(0.16, 0.04)$	$(0.16, 0.04)$	4.00	80.0%
2	2	High	$(0.80, 0.40)$	$(0.64, 0.16)$	4.00	80.0%
2	2	Medium	$(0.60, 0.30)$	$(0.36, 0.09)$	4.00	80.0%
2	2	Low	$(0.40, 0.20)$	$(0.16, 0.04)$	4.00	80.0%
Panel B: Sample composition						
2	2	Low	$(0.40, 0.30)$	$(0.16, 0.09)$	1.78	64.0%
2	1	Low	$(0.20, 0.10)$	$(0.16, 0.09)$	1.78	64.0%

Notes: This table reports the parameter values for each experimental environment in Experiment 2, summarized in Figure 9. Panel A varies the bag compositions (δ_A, δ_B) to generate different likelihood levels while holding the likelihood ratio constant within each ratio condition. Panel B reports the pure-versus-mixed comparison.

C.2 Sample Characteristics

Table C.2. Summary Statistics: Experiment 2

	Full Sample	P	L/P
Number of participants	200	107	93
Demographic information			
Age	41.3	42.6	39.9
Male	54.0	51.4	57.0
White	69.5	70.1	68.8
English first language	92.5	89.7	95.7
Current student	7.5	8.4	6.5
Full-time employed	38.0	38.3	37.6
Born in U.S.	88.0	86.0	90.3
Prior Prolific approvals	4334.1	4471.2	4176.4
Time taken			
Total time taken (minutes)	18.2	15.4	21.3
Average round time (seconds)	29.7	22.9	37.6
Comprehension checks			
Drawing with replacement correct	94.5	92.5	96.8
Drawing order correct	89.5	91.6	87.1
Sample Chance task correct	78.5	–	78.5
Which Bag task correct	84.0	88.8	78.5

Notes: The table reports sample means for the full sample and by treatment assignment. Percentages are reported for binary variables. Average round time is calculated across the 20 experimental rounds. The Sample Chance comprehension check was administered only in the L/P treatment and is therefore blank for the Posterior treatment.

C.3 Likelihood Judgments by Number of Signals

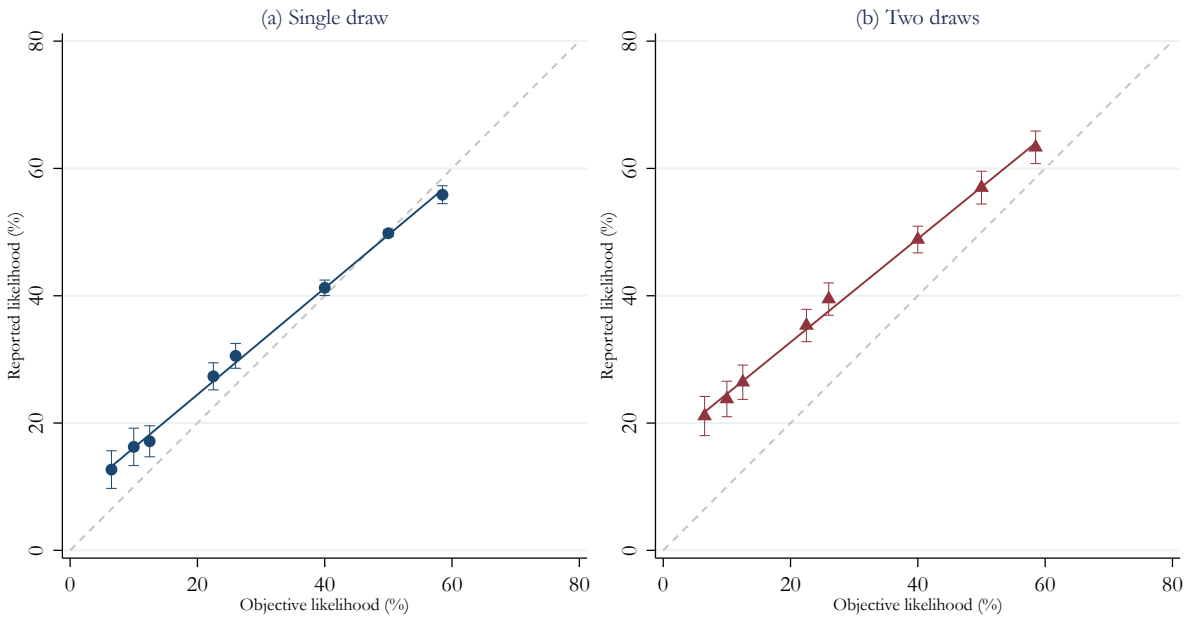


Figure C.1. Reported versus Objective Likelihoods by Number of Signals, Experiment 2

Notes: Panel (a) shows the eight single-draw environments ($M = 1$), while panel (b) shows the ten two-draw environments ($M = 2$), including both pure and mixed samples. Within each participant and environment, reported and objective likelihoods are first averaged across Bag A and Bag B. Points denote mean reported likelihoods across participants, error bars represent 95% confidence intervals, the solid line denotes the best linear fit, and the dashed line denotes the 45-degree identity line.

C.4 Posterior Beliefs by Number of Signals

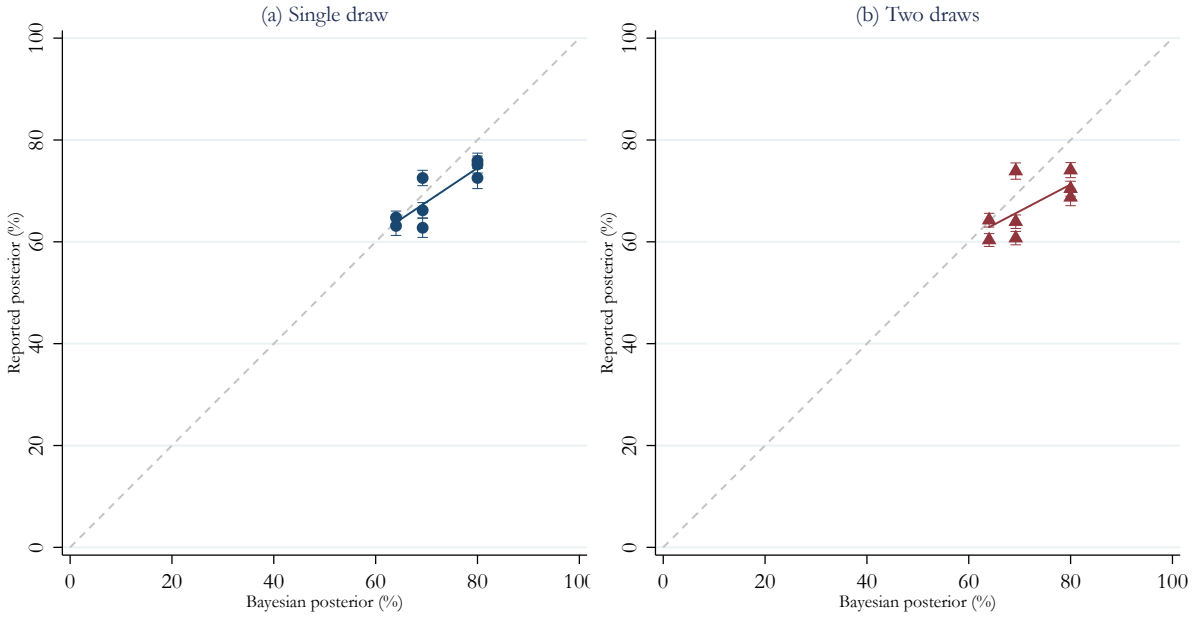


Figure C.2. Reported versus Bayesian Posteriors by Number of Signals, Experiment 2

Notes: Panel (a) shows the eight single-draw environments ($M = 1$), while panel (b) shows the ten two-draw environments ($M = 2$), including both pure and mixed samples. Points denote mean reported posteriors across participants, error bars represent 95% confidence intervals, the solid line denotes the best linear fit, and the dashed line denotes the 45-degree identity line.

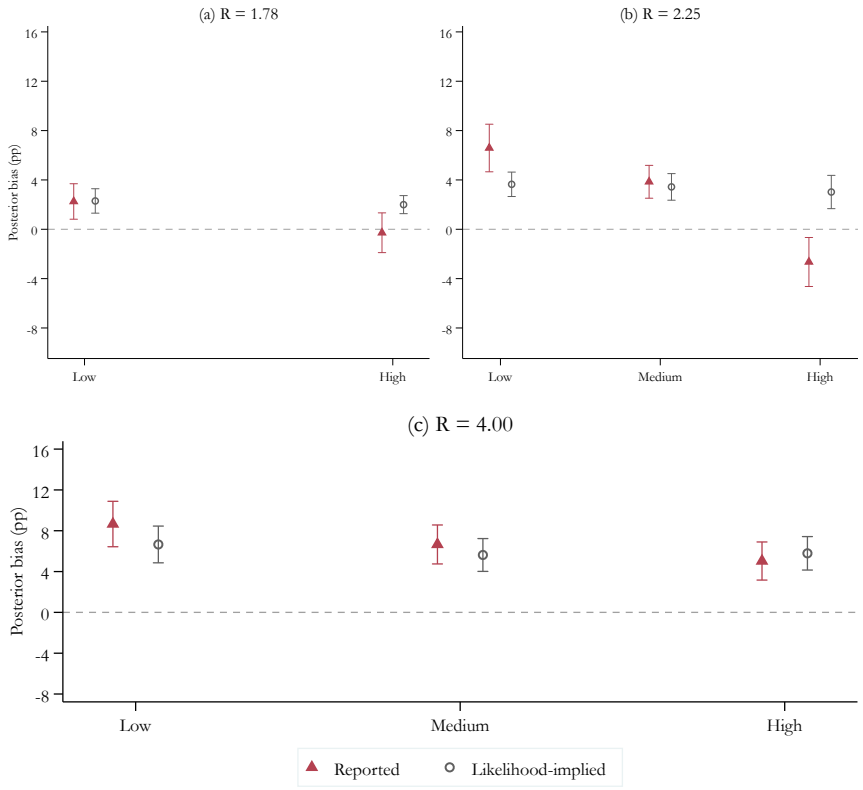


Figure C.3. Reported and Likelihood-Implied Posterior Bias, L/P Treatment

Notes: Each panel corresponds to one objective likelihood-ratio condition $R = L_A/L_B$ and compares low-, medium- (where available), and high-likelihood environments from Panel A of Figure 9 for participants in L/P. Points show mean reported minus Bayesian posteriors, in percentage points, separately for reported and likelihood-implied posteriors, so positive values indicate under-inference. Error bars denote 95% confidence intervals; the dashed horizontal line indicates zero bias.

C.5 Qualitative Evidence Methodology: Experiment 2

As in Experiment 1, participants were asked at the end of the experiment to briefly describe how they formed their probability judgments. Participants in the P treatment explained how they estimated posterior probabilities, while participants in the L/P treatment answered two separate questions about how they formed their sample likelihood and posterior judgments.

Following the approach described in Section B.4, the open-ended responses were reviewed using an LLM to identify recurring reasoning patterns across treatments. The analysis employed the same coding framework as Experiment 1 to maintain comparability, with responses coded using binary, non-exclusive indicators based on explicitly stated reasoning rather than inferred computational strategies. For participants in L/P, the two responses were examined separately to distinguish reasoning about likelihood construction from reasoning about posterior formation.

The resulting qualitative themes and the prevalence of each reasoning feature are reported in Appendix Tables C.3 and C.4, respectively. These descriptions provide complementary evidence on how participants approach the two stages of belief updating and whether the reasoning patterns documented in Experiment 1 persist under the modified experimental parameters.

Treatment	**Common patterns in open-ended responses**
P	Predominantly proportion-based reasoning, comparing the colors in the observed sample with their relative frequencies in each bag. Participants frequently assigned higher probabilities to the bag containing more balls of the observed color, adjusting estimates based on differences between the bags and the number of draws. Some explicitly used likelihood ratios or normalization, while others relied on intuition or approximate calculations.
L/P	More structured two-stage reasoning. In Sample Chance questions, many participants described using the stated probabilities for single draws and multiplying probabilities across multiple draws. In Which Bag questions, participants frequently described comparing the resulting likelihoods and normalizing them to obtain posterior probabilities. Others relied on relative comparisons, proportional adjustments, or intuitive estimates, particularly when converting likelihoods into posterior judgments.

Table C.3. Qualitative themes in open-ended responses by treatment: Experiment 2.

	P	L/P
Proportion-based reasoning	68.9%	51.6%
Multiplicative likelihoods	1.9%	35.5%
Conditional likelihood framing	20.8%	39.8%
Explicit prior mention (50/50)	3.8%	6.5%
Explicit normalization (ratios)	11.3%	24.7%
Prior–likelihood integration	3.8%	6.5%
50/50 fallback	1.9%	4.3%
Guessing / intuition language	33.0%	40.9%
Averaging / add-then-divide error	2.8%	5.4%

Table C.4. Reasoning patterns in open-ended responses by treatment: Experiment 2.
Notes: Each participant's open-ended responses are coded for the presence of non-exclusive reasoning features using the coding framework from Experiment 1. For L/P, a feature is coded as present if mentioned in either the Sample Chance or Which Bag response. Entries report the percentage of valid responses within each treatment mentioning the corresponding feature ($N_P = 106$, $N_{L/P} = 93$). Coding is based on stated reasoning rather than inferred numerical strategies.

Table C.4 reveals patterns broadly consistent with Experiment 1. Participants in P predominantly describe proportion-based reasoning (68.9%), comparing the composition of the observed sample with the relative frequencies of colors in each bag. In contrast, participants in L/P are substantially more likely to report explicit multiplicative likelihood construction (35.5% versus 1.9%), conditional likelihood comparisons (39.8% versus 20.8%), and normalization of likelihoods into posterior probabilities (24.7% versus 11.3%). However, references to guessing or intuition remain common in both treatments (33.0% in P and 40.9% in L/P), suggesting that eliciting likelihoods encourages more structured computational reasoning without eliminating reliance on heuristics. Together, these descriptions reinforce the findings from Experiment 1 that the judgments participants are asked to make are associated with systematic differences in their reported reasoning strategies.

D Theoretical Derivations

D.1 An Affine Model of Likelihood Distortions

This section illustrates how an additive upward distortion in likelihood judgments can generate posterior under-inference even when subjective likelihoods are combined correctly through Bayes' rule. The model also yields a prediction for how posterior attenuation varies with absolute likelihood levels at a fixed objective likelihood ratio.

Setup. Consider objective likelihoods $L_A, L_B \in [0,1]$ with $S \equiv L_A + L_B > 0$ and equal priors. The Bayesian posterior for Bag A is $P = L_A/S$. Suppose subjective likelihoods are given by

$$\widetilde{L}_b = \alpha + \beta L_b, \qquad b \in \{A, B\}, \tag{22}$$

where $\alpha \geq 0$, $\beta > 0$, and $\widetilde{L}_b \in [0,1]$ over the environments considered. The parameters are common across bags and held fixed across environments. The intercept α captures an additive upward component, while β governs responsiveness to objective likelihoods. Overall reporting error is $\widetilde{L}_b - L_b = \alpha + (\beta - 1)L_b$. Under conditionally Bayesian posterior formation, the implied posterior is

$$\widetilde{P} = \frac{\widetilde{L}_A}{\widetilde{L}_A + \widetilde{L}_B} = \frac{\alpha + \beta L_A}{2\alpha + \beta S}.$$

Proposition 3. *Under Equation* (22)*, the following statements hold.*

1. ***Posterior attenuation.*** *The implied posterior satisfies*

$$\widetilde{P} - \tfrac{1}{2} = \kappa(S)\big(P - \tfrac{1}{2}\big), \qquad \kappa(S) = \frac{\beta S}{2\alpha + \beta S}. \tag{23}$$

   If $\alpha > 0$, then $\kappa(S) \in (0,1)$: for $P \neq 1/2$, $\widetilde{P}$ lies strictly between $1/2$ and P, while $P = 1/2$ implies $\widetilde{P} = 1/2$. If $\alpha = 0$, then $\widetilde{P} = P$.

2. ***Likelihood-level dependence.*** *If $\alpha > 0$, then $\kappa(S)$ is strictly increasing in S. Consequently, any admissible proportional increase $(L_A, L_B) \mapsto (cL_A, cL_B)$ with $c > 1$ leaves the Bayesian posterior unchanged but moves the implied posterior strictly farther from $1/2$ whenever $P \neq 1/2$.*

3. ***Compression of relative evidence.*** *If $\alpha > 0$ and $L_A > L_B > 0$, let $r = L_A/L_B > 1$. Then*

$$1 < \frac{\widetilde{L}_A}{\widetilde{L}_B} = \frac{\alpha + \beta r L_B}{\alpha + \beta L_B} < r.$$

Holding r fixed, the subjective likelihood ratio is strictly increasing in L_B. The corresponding result holds after exchanging the state labels when $L_B > L_A > 0$.

Proof. For part (i),

$$\widetilde{P} - \tfrac{1}{2} = \frac{\beta(L_A - L_B)}{2(2\alpha + \beta S)} = \frac{\beta S}{2\alpha + \beta S}\left(\frac{L_A - L_B}{2S}\right) = \kappa(S)\big(P - \tfrac{1}{2}\big).$$

Since $\beta, S > 0$, $\kappa(S) \in (0,1)$ when $\alpha > 0$, and $\kappa(S) = 1$ when $\alpha = 0$. The directional statements follow immediately.

For part (ii),

$$\kappa'(S) = \frac{2\alpha\beta}{(2\alpha + \beta S)^2} > 0$$

when $\alpha > 0$. Proportional scaling multiplies S by c while preserving P, so the result follows from Equation (23), provided the scaled objective and subjective likelihoods remain in $[0,1]$.

For part (iii), write $\widetilde{r} = (\alpha + \beta r L_B)/(\alpha + \beta L_B)$. Then

$$\widetilde{r} - 1 = \frac{\beta L_B(r-1)}{\alpha + \beta L_B} > 0, \qquad r - \widetilde{r} = \frac{\alpha(r-1)}{\alpha + \beta L_B} > 0.$$

Holding r fixed,

$$\frac{\partial \widetilde{r}}{\partial L_B} = \frac{\alpha\beta(r-1)}{(\alpha + \beta L_B)^2} > 0.$$

□

The attenuation result does not require $\beta < 1$: a positive additive component is sufficient. Holding $\alpha > 0$ and S fixed, a smaller β strengthens attenuation. A purely proportional distortion, by contrast, leaves the implied posterior unchanged because the common factor cancels from the likelihood ratio.

The model illustrates why absolute likelihood levels can affect subjective relative evidence and posterior beliefs even when the objective likelihood ratio is fixed. It motivates the likelihood-level comparison in Experiment 2 without imposing the affine form on the empirical analysis. It does not account for all observed behavior: with $\alpha > 0$, conditionally Bayesian posteriors remain between the prior and the Bayesian benchmark, so the model cannot generate the over-inference observed in some environments.

E Experimental Materials

E.1 Experiment 1

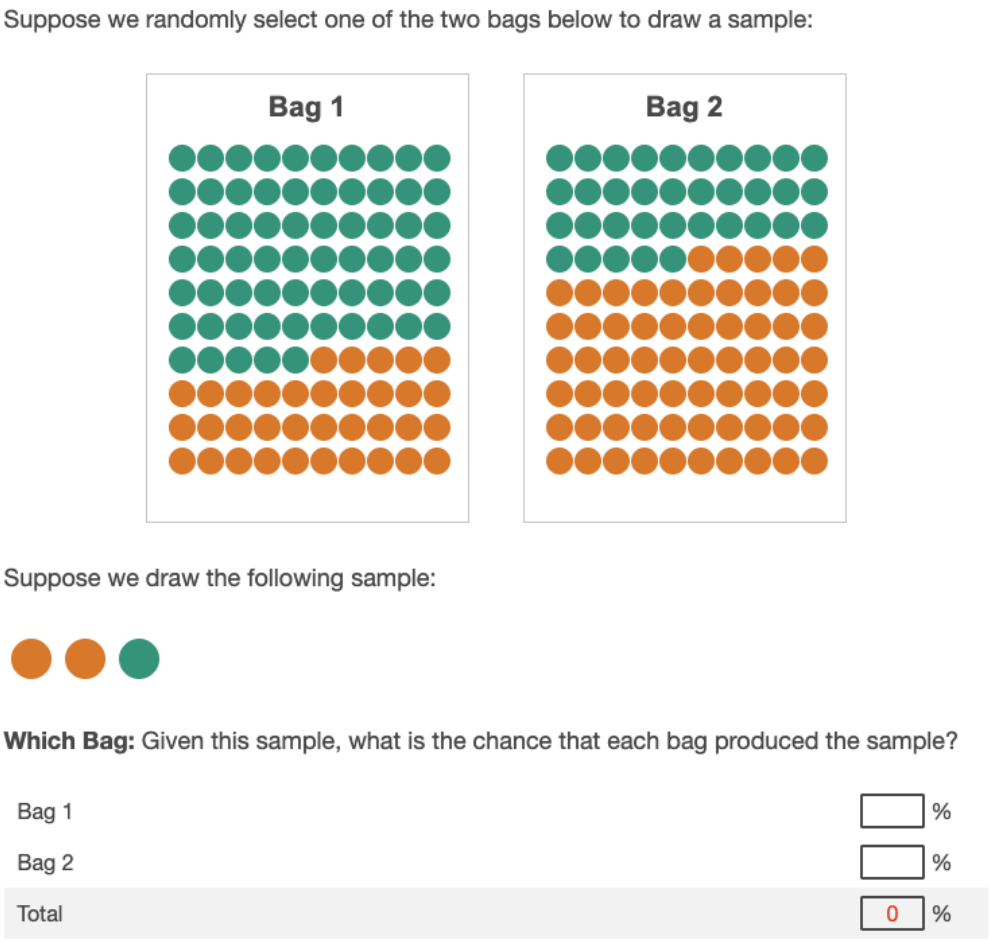


Figure E.1. Experimental task screen in the P treatment, where participants reported posteriors only

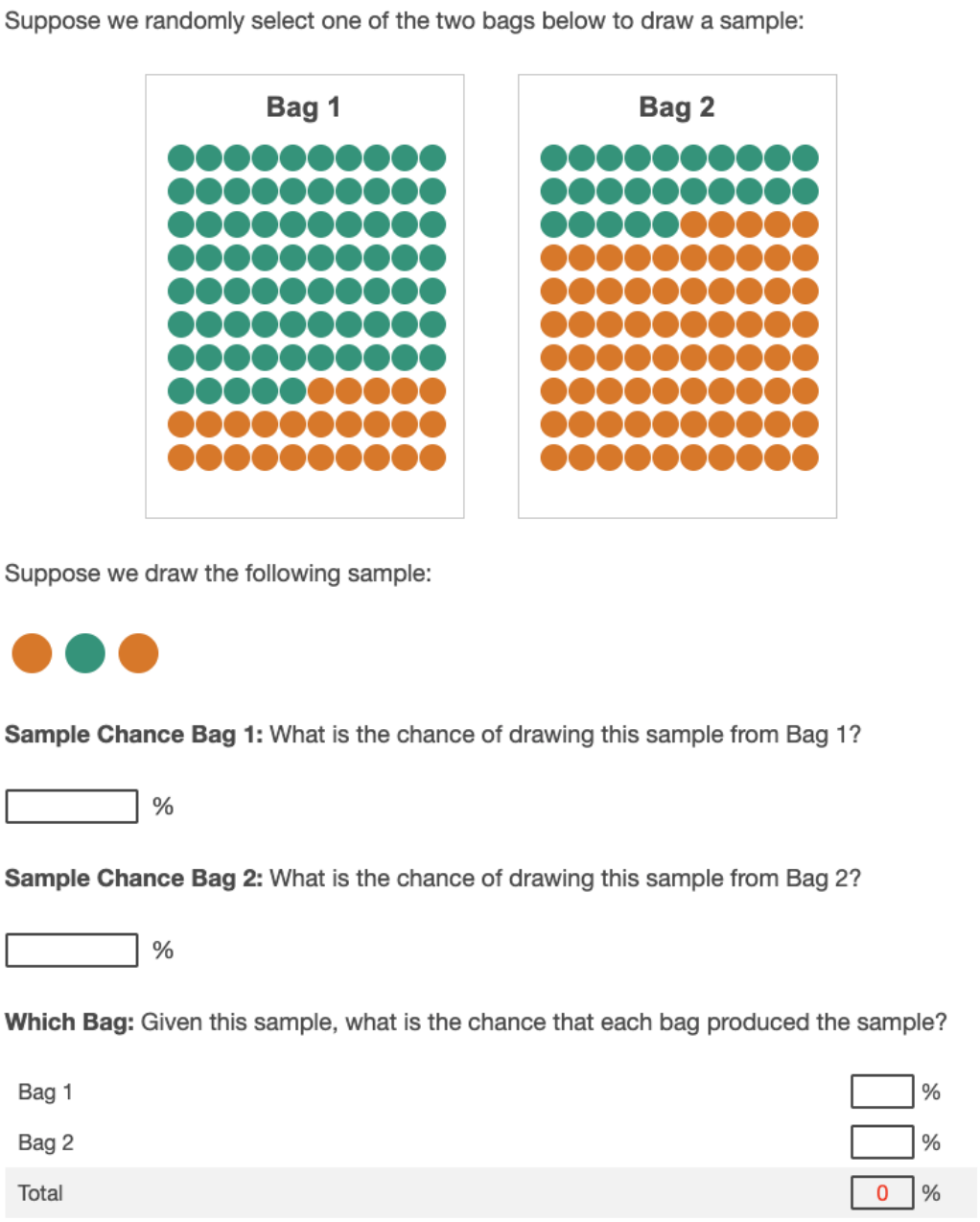


Figure E.2. Experimental task screen in the L/P treatment. Participants reported both likelihoods and posteriors

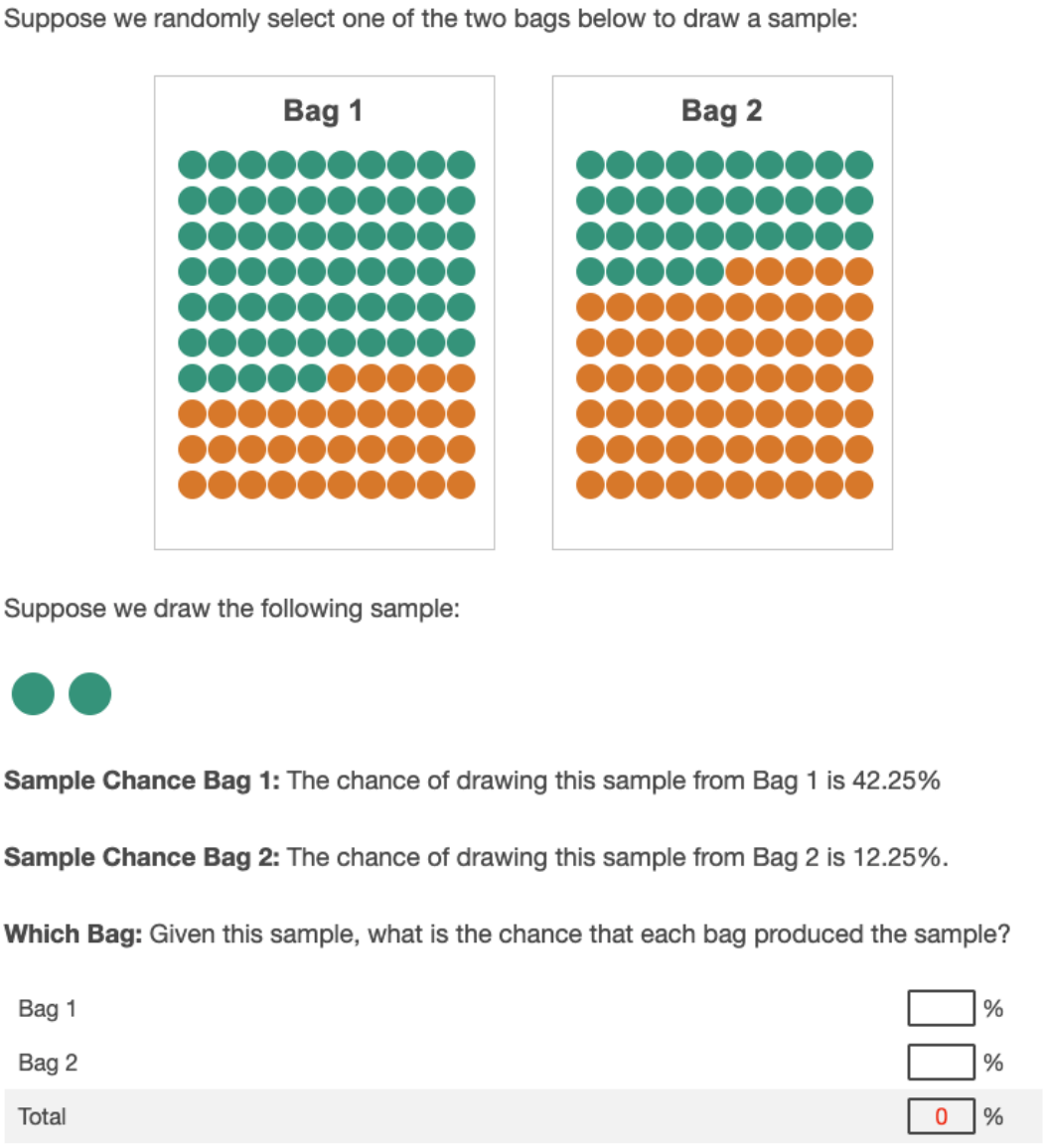


Figure E.3. Experimental task screen in the L*/P treatment. Participants were provided likelihoods for each bag and reported posteriors

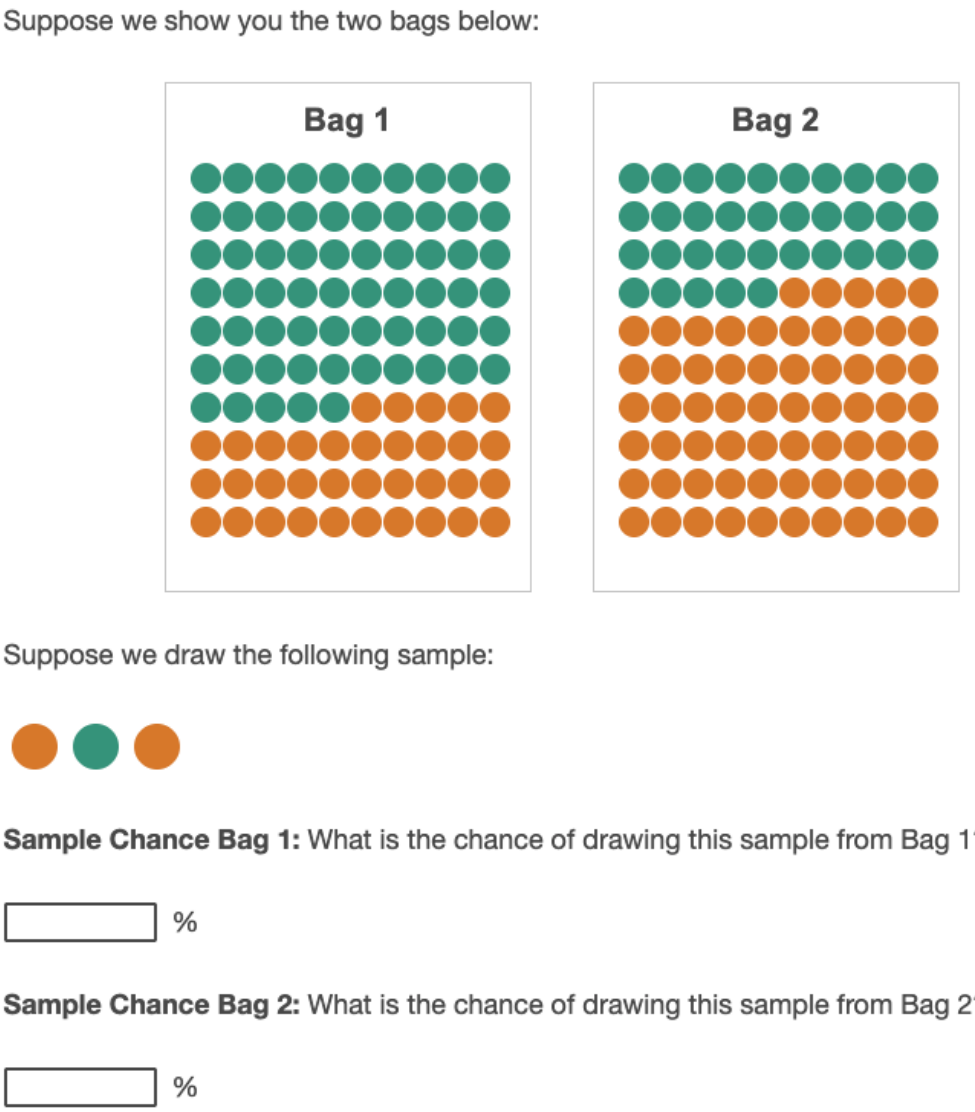


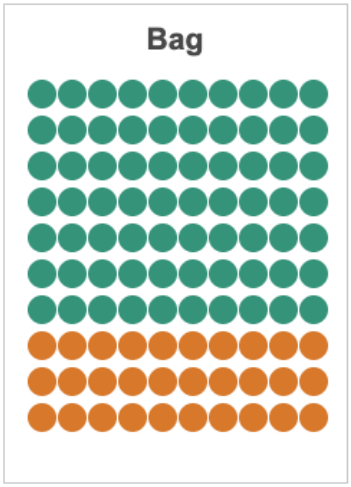
Figure E.4. Experimental task screen in the L treatment, where participants reported likelihoods only

Overview

This survey has a total of 15 problems. Each problem has multiple questions.

Each question has a correct statistical answer, which will determine your bonus payment if it is selected.

Let's start with an example. Here is a bag filled with 70 Green and 30 Orange balls:



Click the continue button to see how we'll use this bag in the problems.

How We Draw Samples

Here's a **sample of balls** drawn **in order** from the bag you just saw:



This is how we drew the above sample of balls from the bag:

1. We picked up the bag and drew one ball at random. It was Green.
2. We dropped the ball back into the bag, so the mix of balls stayed the same, shook the bag, and drew a second ball. It was Orange.
3. We repeated the process one more time and drew a third ball. It was Orange.

When we show you a sample of balls, it will be drawn this way. You'll see how many balls are Green and how many are Orange.

Click on the continue button to answer some questions to make sure you understood how we draw samples.

Figure E.5. Drawing with replacement instructions for all treatment groups

We will now ask you some questions to see if you have understood how we draw samples.



Suppose we draw the following sample from the bag above:



When we draw balls from the bag, what happens after each ball is drawn?

The ball is put back into the bag, the mix stays the same, and the bag is shaken before drawing again

The ball is kept out of the bag, so the mix changes for the next draw

You know the exact order the balls were drawn — 2 Green balls first, then 1 Orange ball.

True

False

Figure E.6. Sampling comprehension check for all treatment groups

Sample Chance Questions

Let's start with an example. Below are two bags:



Suppose we show you the following sample:



If we draw 3 balls in order from Bag 1, there is a correct statistical chance of getting the exact sample shown above.

If we instead draw 3 balls in order from Bag 2, there is also a correct statistical chance of getting this sample.

Note that these numbers **do not need to add to 100** — we're judging them separately.

In Sample Chance questions, **your task is to estimate the chance of drawing the exact sample from each bag** by entering a number between 0 and 100:

1. **Sample Chance Bag 1:** If we draw 3 balls from Bag 1, what is the chance of drawing the above sample?
2. **Sample Chance Bag 2:** If we draw 3 balls from Bag 2, what is the chance of drawing the above sample?

Click the continue button below to learn about Which Bag questions.

Figure E.7. Sample Chance Instructions Screen, L/P Treatment

Which Bag Questions

Let's look at another example. Below are two bags:



Below is a sample of balls:



Suppose we randomly select one of the two bags above to draw this sample.

Given this sample, there are correct statistical chances that the bag we are using to draw the sample is either Bag 1 or Bag 2.

In Which Bag questions, **your task is to estimate the chance that each bag produced the sample** by entering a number between 0 and 100 for:

- Bag 1
- Bag 2

These chances **must add up to 100 percent** — if a sample didn't come from one bag, it must have come from the other.

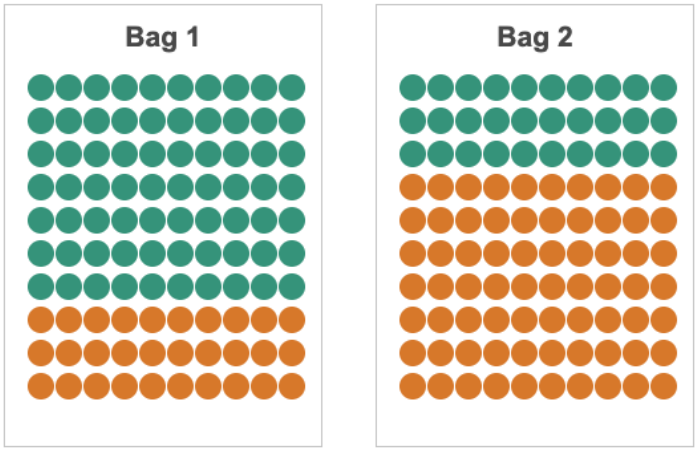
Each problem is independent of the previous ones — a new bag and sample may be chosen for each.

Click the continue button below to learn more about how you'll be paid.

Figure E.8. Which Bag Instructions Screen, L/P Treatment

We will now ask you some questions to see if you have understood the instructions.

Suppose we randomly select one of the two bags below to draw a sample:



Suppose we draw the following sample:



What are you estimating in a Sample Chance question?

- How many balls are in each bag
- The chance that each bag produced the sample
- The chance of drawing this exact sample from each bag

What are you estimating in a Which Bag question?

- How many balls are in each bag
- The chance of drawing this exact sample from each bag
- The chance that each bag produced the sample

Figure E.9. Probability-task comprehension check (L/P treatment) Feedback was provided after submission.

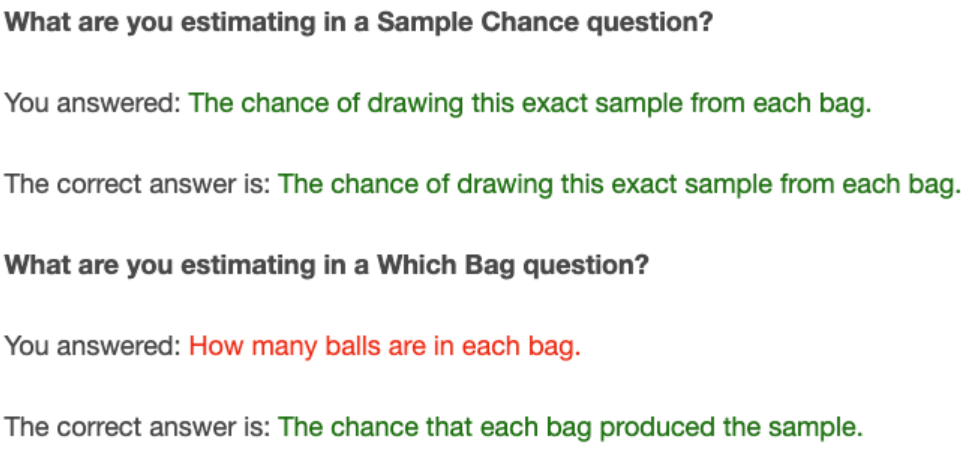


Figure E.10. Task comprehension check feedback screen (L/P treatment)

Notes: After submitting their answers, participants in L/P were shown their responses alongside the correct answers. Incorrect responses were highlighted in red and correct responses in green, reinforcing the distinction between likelihood ("Sample Chance") and posterior ("Which Bag") probabilities before proceeding to the main task.

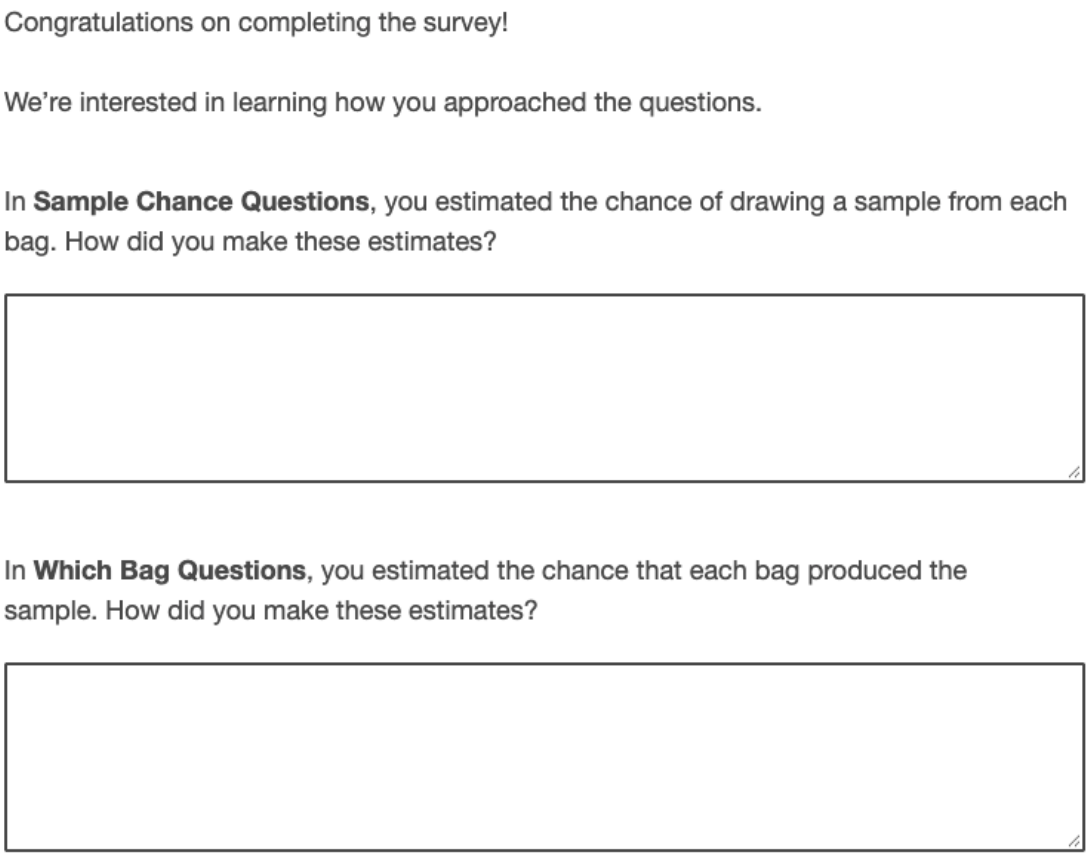


Figure E.11. Open Ended Response Screen, L/P Treatment

Earnings

At the end of the study, we'll randomly select one out of every five participants to receive a bonus payment.

If you are selected, we'll randomly pick one of the 15 problems you answered.

From that problem, we'll randomly pick one of the three questions you answered: two Sample Chance questions and one Which Bag question.

You'll earn $10 if your answer on that question is within **5%** of the correct statistical answer.

Any of the 15 problems could be selected, so treat each problem carefully as if it's the only one that counts.

Click on the continue button below to answer some questions and make sure you understood the instructions.

Figure E.12. Incentives Screen, L/P Treatment

E.2 Experiment 2

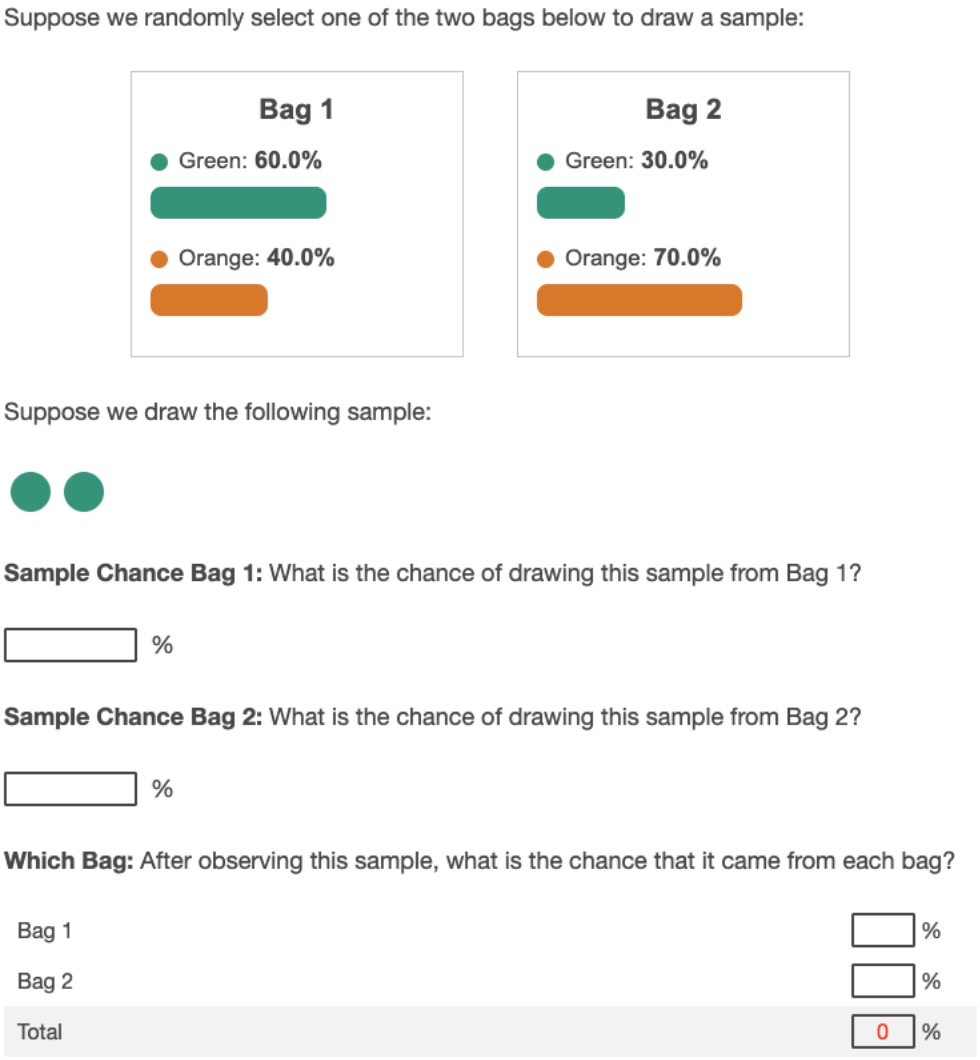


Figure E.13. Experimental Interface for L/P Treatment, Experiment 2

Notes: The interface is identical to that used in Experiment 1 except that the bag compositions are asymmetric and displayed as percentages rather than counts. Participants observe an ordered sample of one or two draws with replacement before reporting likelihood judgments (L/P treatment) and/or posterior beliefs. The experiment labels the bags as "Bag 1" and "Bag 2," corresponding to Bags A and B in the text.

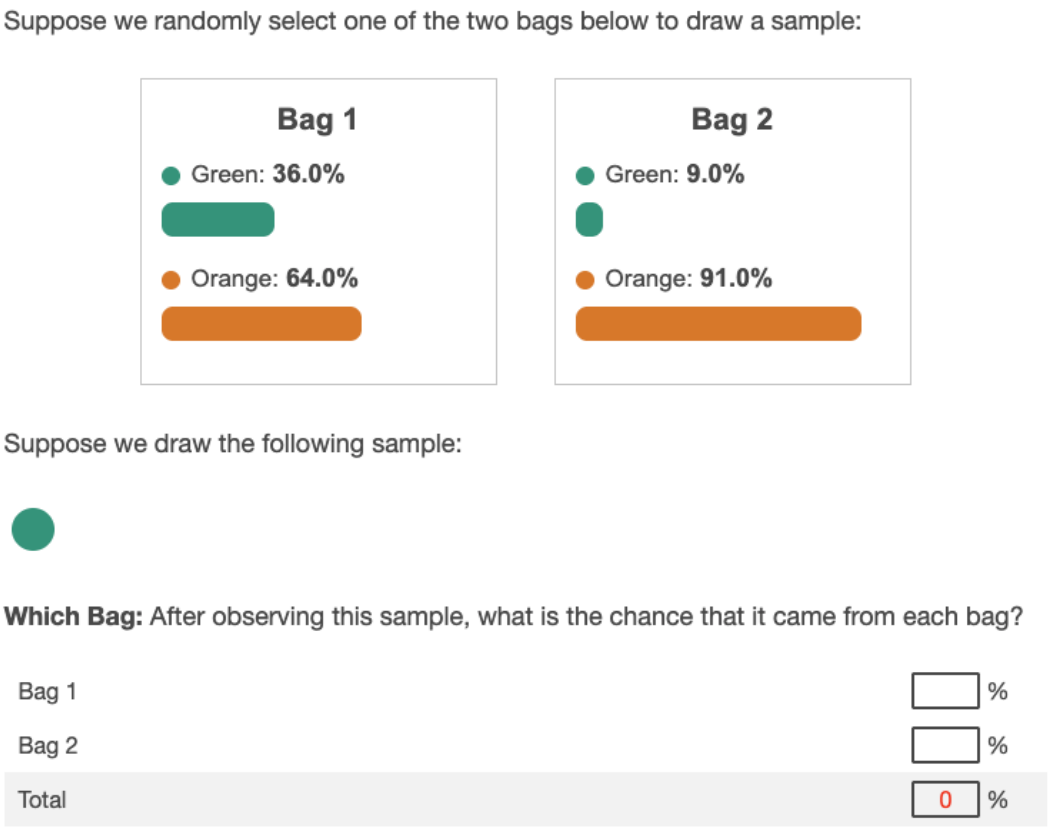


Figure E.14. Experimental Interface for P Treatment, Experiment 2

Notes: The interface is identical to that used in Experiment 1 except that the bag compositions are asymmetric and displayed as percentages rather than counts. Participants observe an ordered sample of one or two draws with replacement before reporting posterior beliefs. The experiment labels the bags as "Bag 1" and "Bag 2," corresponding to Bags A and B in the text.

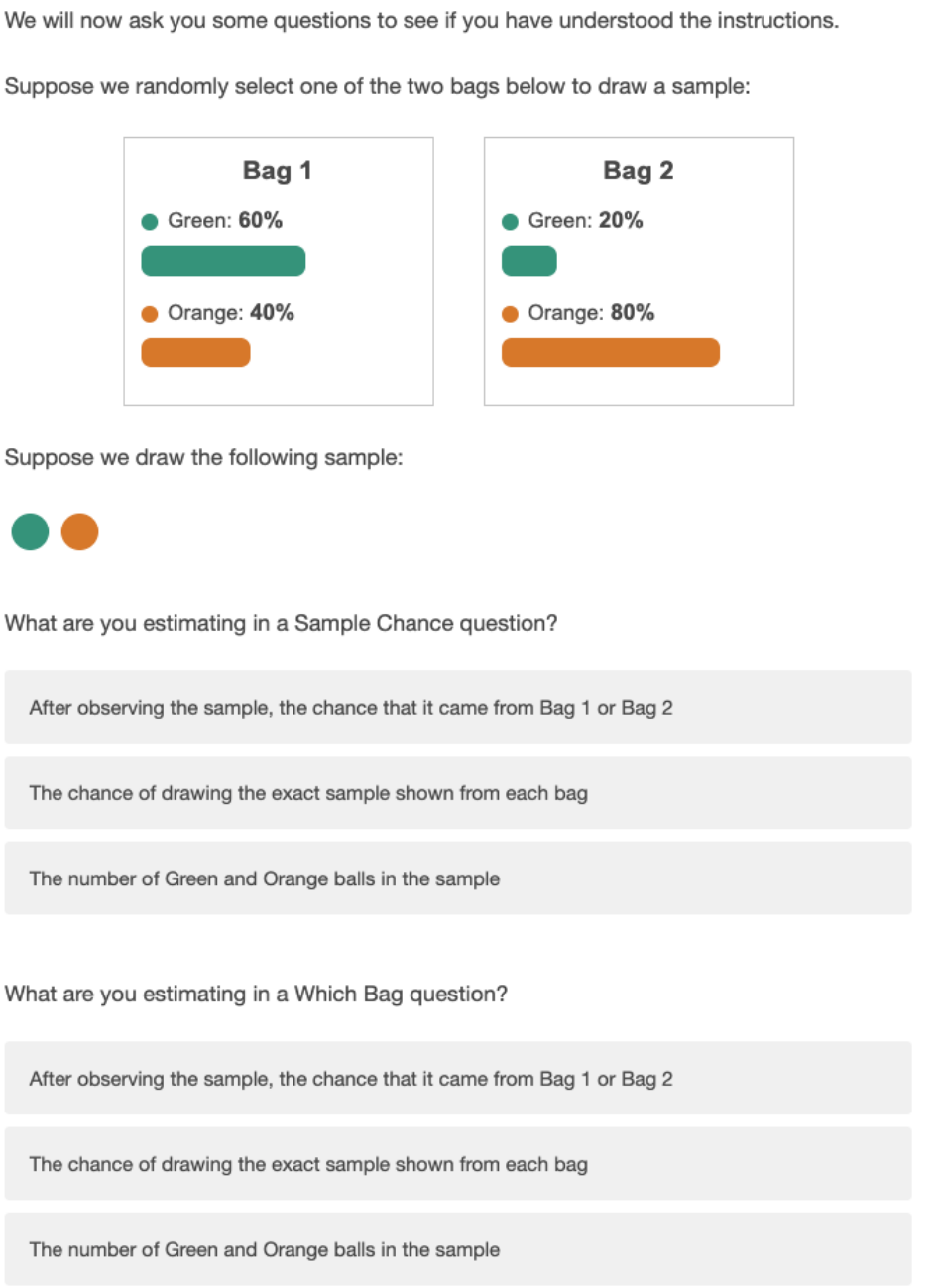


Figure E.15. Probability-task comprehension check (L/P treatment, Experiment 2). Feedback was provided after submission.

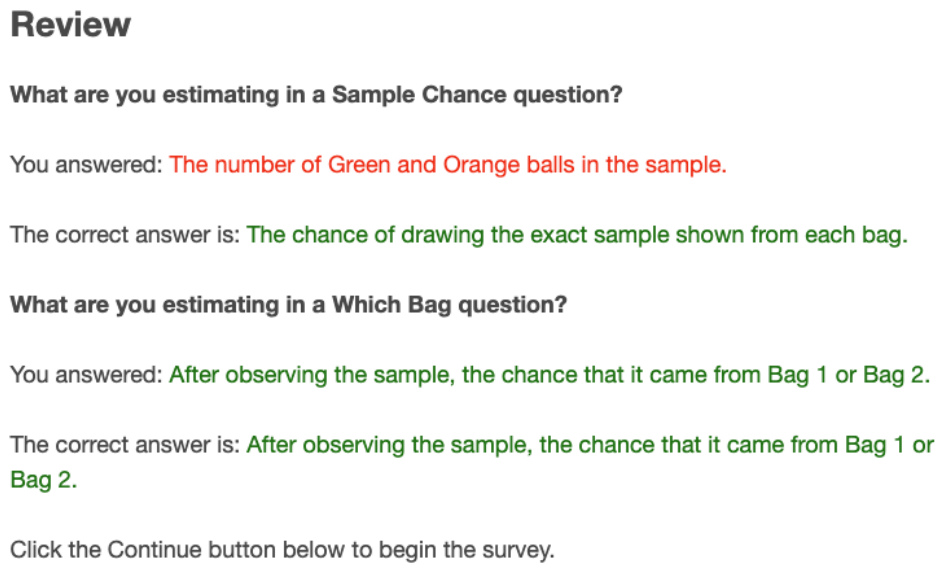


Figure E.16. Task comprehension check feedback screen (L/P treatment), Experiment 2

Notes: After submitting their answers, participants in L/P were shown their responses alongside the correct answers. Incorrect responses were highlighted in red and correct responses in green, reinforcing the distinction between likelihood ("Sample Chance") and posterior ("Which Bag") probabilities before proceeding to the main task.