Confidence in Inference*

En Hua Hu
University of Toronto

March 7, 2024

Abstract

I study a decision-maker who chooses between objects, each associated with a sample of signals. I axiomatically characterize the set of choices that are consistent with established models of belief updating. A simple thought experiment yields a natural choice pattern that lies outside this set. In particular, the effect of increasing sample size on choice cannot be rationalized by these models. In a controlled experiment, 95% of subjects' choices violate models of belief updating. Using a novel incentive-compatible confidence elicitation mechanism, I find confidence in correctly interpreting samples influences choice. As suggested by the thought experiment, many subjects display a sample size neglect bias which is positively associated with higher confidence.

keywords: Ambiguity, belief updating, confidence, imprecise cognition, sampling

JEL codes: C91, D81, D83, D91

To view the latest version of the paper: Latest Version

*I am deeply grateful to my supervisors and committee, Yoram Halevy, Colin Stewart, and Marcin Peski, for their unwavering support and guidance. I extend particular thanks to Gabriel Caroll, Yucheng Liang, and David Walker-Jones for their insightful discussions and feedback. My appreciation also goes to Marina Agranov, Itai Arieli, Heski Bar-Isaac, Rahul Deb, Tanjim Hossain, Stanton Hudja, Rohit Lamba, Giacomo Lanzani, Ilya Segal, Jakub Steiner, Tomasz Strzalecki, Martin Vaeth, Michael Woodford, Leeat Yariv, Lanny Zrill and the audiences of various seminars and conferences for their valuable comments and discussions. Special mention to Stanton Hudja for his exceptional assistance with the logistics of the experiment. The experiment is pre-registered on Aspredicted.org.

1 Introduction

Much of the information used in decision-making comes in the form of a sample of signals. This ranges from comparing different Google Maps reviews before deciding on a restaurant to gathering several referee reports before a verdict. Given the ubiquity of samples in inference, I investigate how they impact choice behavior.

Previous work has focused on documenting non-Bayesian belief patterns and accommodating them via different belief updating models.[1] This paper considers instead choice behavior between objects associated with signals. The main question I investigate is whether models of updating can accommodate the observed choice patterns. The sheer number of non-Bayesian models makes this a difficult query. To address this challenge, Theorem 1 axiomatically identifies the choice behavior of an extremely general class of updating models. This enables direct testing experimentally via revealed preference. I find that 95% of subjects display choice patterns that these models cannot rationalize. My results further hint that the discrepancy lies in these models ruling out the possibility that decision-makers (DM) may lack confidence in correctly interpreting information.

I illustrate the core behavioral violation of these models via a thought experiment. Alice, a venture capitalist, is choosing to invest between two projects, A and B. Each project can either succeed or fail. The outcomes of the projects are independent. Alice considers them equally likely to succeed so she consults experts on these projects. Of the experts consulted for project A, 4 out of 5 predict its success. For project B, 1 out of 1 expert predicts success. Alice assumes the experts' predictions are identically and independently distributed (iid) conditional on the outcome with some known likelihood of correctness. Suppose she later learns additionally that 45 experts have analyzed A, and 36 predict success, making a total of 40 out of 50. For B, 9 additional experts unanimously predict success, making a total of 10 out of 10. A justifiable choice might be to initially pick project A when sample sizes are low, and switch to B when sample sizes are multiplied tenfold. One might of course argue that the correct choices and when to switch depend on the likelihood of prediction correctness. However, it turns out that irrespective of the likelihood's value, this switching pattern cannot occur for a very general class of belief-updating models. I formalize this result theoretically via an axiomatic characterization, experimentally verify the robustness of the thought experiment in a more extensive setting, and suggest additional relevant channels that these models rule out which are important for choice.

My theoretical framework features a DM choosing between ex-ante identical objects for each of which they observe a sample of signals. I consider the class of DM whose updating rule is monotonic in the likelihood ratio of samples, computed under the assumption that signals are iid with known likelihoods. Monotonicity is a weak assumption that holds universally for updating rules, and both theoretical as well as experimental works commonly feature iid signals with known likelihoods. Additionally, I do not restrict the DM to be expected utility maximizing. Theorem 1 shows that the choices of such DMs are axiomatically characterized by a *separability* axiom and

[1] See Benjamin (2019) for a survey.

mild regularity axioms.[2] The separability axiom states that if an object with sample x is chosen over another with sample y, denoted by $x \succ y$, then when any other sample z is added, an object with sample $x + z$ will still be chosen over one with $y + z$. Separability says that adding the same sample to two others does not reverse preference, which contradicts the thought experiment. If $x \succ y$, then separability implies $x + x \succ y + x$, and $y + x \succ y + y$. Transitivity then implies $x + x \succ y + y$. This process applied 10 times gives that $x \succ y$ implies $10 \times x \succ 10 \times y$, ruling out the switching pattern from the thought experiment.

Almost all information processing models feature Bayesian updating in addition to the DM facing no uncertainty regarding the signal likelihoods - falling under Theorem 1.[3] Theorem 1 pins down the empirical content of these standard assumptions in this sampling environment. Theorem 1 additionally enables direct experimental testing of these assumptions through choices instead of having to elicit beliefs. This is crucial, as failure of separability may be due to failure of probabilistic sophistication. In this case, belief elicitation methods are not incentive-compatible and the elicited measures do not have a clear interpretation.

I test separability through a controlled experiment where participants choose between boxes filled with colored balls. Without knowing each box's type (good or bad), which affects ball distribution, subjects make choices based on observed draws, aiming for a bonus for choosing a "good" box. This setup reflects the thought experiment and the theoretical set-up, with projects as boxes and expert predictions as balls. I induce from choices, for each subject, a set of indifference curves. I show that choices satisfying separability must have a linear representation, which is reflected in indifference curves being parallel straight lines. By analyzing these curves' angles and their standard deviations, I test for parallelism, where a near-zero standard deviation indicates separability holds.[4] I find that the average standard deviation of angles is 27 degrees. Furthermore, only 5% of subjects have a standard deviation of less or equal to 10 degrees. These results show that updating models satisfying separability are systematically rejected.

These findings suggest that updating fails due to either failure of monotonicity or subjective uncertainty regarding signal likelihoods. Two streams of literature have proposed the second explanation. Recent works in ambiguity [5] and cognitive imprecision[6] show that subjective uncertainty regarding the data generating process can lead to dynamics very different from the standard paradigm. These works model such uncertainties explicitly and show that choice behavior may be consistent with the predictions of particular models. My results instead reject, given monotonicity, *any* model which *does not* feature this subjective uncertainty. Thus offering evidence for the existence as well as the relevance of this channel without committing to a particular way of modeling this uncertainty. The reader may find it helpful to view my results through the lens of Ellsberg

[2]The mild regularity axioms are completeness, transitivity, and continuity.

[3]The large majority of works in information design, information acquisition, and social learning fall under this category.

[4]Since signal numbers are discrete, it is not necessarily precisely 0.

[5]Epstein and Schneider (2007); Epstein and Halevy (2019, 2023); Ngangoué (2021); Kellner et al. (2022); Liang (2023); Shishkin and Ortoleva (2023)

[6]Steiner and Stewart (2016); Woodford (2020); Khaw et al. (2021); Frydman and Jin (2022); Enke and Graeber (2023)

(1961), which shows that DMs are not probabilistically sophisticated. Ellsberg (1961) is celebrated because it gives this previously elusive mental phenomenon its behavioral and empirical counterpart. I make an analogous argument and give a behavioral pattern that rejects models without uncertainty regarding signal interpretation.

The thought experiment offers insights as to why separability rules out uncertainty in signal interpretation. Consider again the thought experiment: Should Alice be more confident in her first choice of 4 out of 5 over 1 out of 1 or her second choice of 10 out of 10 over 40 out of 50? By "confident", I refer to Alice's confidence in selecting the project with the highest probability of success. This directly ties to her confidence in correctly interpreting signals. One could argue that as the sample sizes grow, Alice learns to interpret predictions better and becomes more confident. Therefore, it might be natural to say that Alice is more confident in her second choice. This translates to being more confident in having interpreted signals correctly in the second choice. However, if signal likelihoods are known and iid, then signal interpretation should be independent of the observed sample. In other words, separability implies a mental model where the DM already knows the signal informativeness and rules out the role of confidence in choice. I further highlight two features of the choice process in the thought experiment. First, Alice chose solely based on sample characteristics: sample sizes and proportions of success. In particular, her choices were made without referring to any information regarding signal likelihoods. Second, as the sample size grows, one is more comfortable neglecting the sample size and choosing by the proportion of success, which occurs in parallel to one's increasing confidence. Suggesting one neglects the sample size when one is confident in having seen a high enough sample size.

Previous works have also found non-Bayesian attitudes towards size and proportion. Griffin and Tversky (1992) finds that subjects underweight the sample size via belief elicitation, they suggest that the bias is constant. The thought experiment suggests that the channel is more nuanced. In particular, the weight given to the sample size is high when the samples have small sizes and decreases as more samples are observed. A model with constant bias, as they suggest, cannot account for the thought experiment.

To test these features of choice and the relevance of confidence, I structured my experiment with three distinct between-subject treatments, each involving a different information structure. This allows me to test whether subjects refer to the objective likelihoods or only to sample sizes and proportions. Additionally, I introduce and implement a novel incentive-compatible confidence elicitation mechanism. This enables me to test whether neglecting the sample size is associated with higher confidence. As per my pre-registration, I run my analysis on the full sample and a sub-sample of subjects who satisfy a weak coherence condition. This sub-sample of subjects displays a greater understanding of the experiment, allowing me to check the robustness of results and ensure results are not driven by confusion and inattention.

My three treatments differ in the information structure provided to subjects. Two of the three information structures have iid signals with known but different likelihoods, while the third information structure features uncertain signal likelihoods. Subjects are told explicitly about these information structures. These differences allow me to test whether the likelihood matters and whether subjects ignore the likelihood and choose entirely based on sample characteristics. I find that the information structure has virtually no effect on the subjects' choices - almost all subjects violate separability, and their choices are identical under all three treatments. This aligns with the thought experiment and suggests that subjects ignore the information structure and choose based on sample characteristics instead. When looking at the sub-sample, I find the same result, confirming that this is not driven by confusion or the complexity of the environment but rather the outcome of intentional choice.

Experimental work in belief updating has traditionally given subjects the objective information structure (Grether (1980); Coutts (2019); Barron (2021); Möbius et al. (2022)). Belief updating biases are typically then estimated *assuming* that subjects form a correct subjective mental model of this objective information structure. In this sampling environment, subjects are strongly insensitive to the objectively given information structures. This suggests the link between objective and subjective structures is likely much looser than previously thought. This has important implications as Shmaya and Yariv (2016) shows that even Bayesian updating can seldom be rejected without assumptions on the mental subjective information structure. Similarly, my results show that given the standard assumptions of these models, monotonicity, and iid signals with known likelihoods, *every* updating model is rejected. Taken together, these results suggest that while the literature has focused on updating rules, we need to understand better the subjective mental model of the information structure to explain choice behavior.

If subjects' choice is influenced by potential uncertainty in signal interpretation, then this implies they might not be confident in their beliefs. To study this channel, one needs to measure confidence in beliefs, a second-order epistemic object. Several strands of the literature attempt to measure confidence: ambiguity, incomplete preferences, dynamic beliefs, and cognitive imprecision. The elicitation methods proposed are however typically either not incentive-compatible (Enke and Graeber (2023); Nielsen and Rigotti (2023)), or only incentive-compatible under strong assumptions (Karni (2018, 2020); Chambers and Lambert (2021)), or require restrictive environments (Coffman (2014); Halevy et al. (2023)).[7] I propose a confidence elicitation method that is simple for subjects to understand, incentive-compatible for a large class of theories, and also has low implementation cost. I define confidence as knowledge of the correct action. This mental phenomenon, under general theories, is shown to be tied behaviorally to the willingness to pay for information. This allows me to measure the lack of confidence by the willingness to pay to learn the correct action. I show that asking just one additional and simple-to-understand binary choice question after any standard choice or belief elicitation task is sufficient. The only requirement is that there is a correct choice (subjective or objective) given the subject's information. This is a very mild requirement, as confidence is typically measured as being confident in having made the correct choice. This measure is also shown to be highly correlated to an unincentivized measure.

The thought experiment suggests that *sample size neglect*, defined as choosing entirely based

[7]The issue has also been explored empirically, see Giustinelli et al. (2022) and Kerwin and Pandey (2023).

on sample proportion, is a sign of confidence. I first document that 39% and 61% of choices in my full sample and sub-sample display sample size neglect, respectively. This is in line with the intuition of the thought experiment, as the sub-sample is shown to be more confident and more likely to neglect the sample size. Moreover, I examine whether displaying sample size neglect is correlated with the choice to incur costly learning. I find a subject is 1.55 times more likely to incur costly learning on choices that do not display sample size neglect. The effect is sharper for the subsample and amongst choices involving larger samples, with subjects being 2.4 times more likely to incur learning when not displaying sample size neglect. Therefore, as suggested by the thought experiment, displaying sample size neglect is associated with being less likely to opt to learn and higher confidence.

I also offer a theoretical foundation for the observed behaviors and the channels documented in the thought experiment. In the real world, DMs frequently encounter uncertainty regarding signal likelihood. For example, one may be uncertain about the harshness of reviewers or the accuracy of experts. As this likelihood is unknown but remains fixed as more signals are gathered, it is possible to learn about it from samples. Therefore, a DM who observes only a few signals is more uncertain, and hence less confident, of her posterior belief. The DM can update her belief about this uncertainty, and it dissipates as the sample size grows. This is reflected by a higher confidence when facing large samples. However, on many occasions, the DM does not even know how to update or form beliefs about the uncertainty. I show that, in this case, under a mild monotonicity condition, no matter what the uncertainty is, sample size neglect is asymptotically optimal. Hence giving a plausible explanation as to why subjects can confidently ignore the likelihoods and choose based on sample proportion.

Organization. The rest of the paper is organized as follows. Section 2 presents a thought experiment that illustrates an intuitive behavior that conflicts with conventional models. Section 3 establishes the environment, the axioms, and the representation result. I also present the confidence elicitation mechanism in Section 3. Section 4 gives an exposition of the experimental design. Section 5 presents the experimental findings. Section 6 shows likelihood uncertainty can accommodate the observed behaviors. Section 7 concludes the paper.

2 Thought Experiment

Alice, a venture capitalist, has two potential projects, A and B that she can invest in. She has only enough funds to invest in one of them. Both projects promise that they can succeed in creating an industry-leading technology. The technologies are from different fields. Therefore, the success of one project is independent of the other. Ex-ante Alice believes both are equally likely to succeed and only cares about whether they succeed. To make a better decision, Alice reaches out to experts in these fields. Experts give out predictions for whether a project will succeed. Alice can assume these experts are predicting independently without any hidden agenda. Therefore, signals are iid conditional on the outcome of the projects. For project A, 4 out of 5 experts predict it will succeed. For project B, only one expert has gotten back to Alice, but they predict success. How should Alice choose? A natural and justifiable choice would be project A1, as a single expert's prediction for B may be deemed insufficient. Now suppose Alice observes some additional signals later. Project A now has 40 out of 50 experts predicting its success, and B now has 10 out of 10 experts predicting its success. Should Alice now be willing to switch to investing in B? If not, what about 400 out of 500 versus 100 out of 100? It may seem natural once the sample sizes grow enough, that Alice should be comfortable with investing in B. Furthermore, should Alice be more confident in the correctness of the first choice or the second? By correct, I mean not in selecting a successful project but having chosen the project with the highest probability of success given the predictions. I suspect one may find it acceptable to be more confident with the second choice. And introspection suggests that as the sample sizes grow, our willingness to focus on the sample proportion *and* our confidence both increase.

If Alice did choose A initially and B later on, then her behavior is inconsistent with a large and general class of models. In the following, I illustrate that a Bayesian EU DM cannot generate such behavior. The theory section shows it holds for a much broader class of DMs. Suppose Alice believes that projects will succeed with probability $p \in (0, 1)$; recall ex-ante Alice considers them equally likely to succeed. Suppose when a project, A or B, does succeed; Alice believes each expert has a probability c_a for project A and c_b for project B of correctly predicting success. When a project does fail, this probability, which is now a false positive, is d_a for project A and d_b for project B. If Alice is Bayesian, Alice will choose whichever project has a higher posterior probability of success given the observed sample of opinions. Then one can derive the condition for Alice to prefer A1 over B1 in terms of posteriors that

$$\Pr(\text{A succeeds} \mid \text{4 out of 5}) > \Pr(\text{B succeeds} \mid \text{1 out of 1}),$$

and because states are binary, the following regarding posterior ratios holds

$$\frac{\Pr(\text{A succeeds} \mid \text{4 out of 5})}{\Pr(\text{A fails} \mid \text{4 out of 5})} > \frac{\Pr(\text{B succeeds} \mid \text{1 out of 1})}{\Pr(\text{B fails} \mid \text{1 out of 1})}.$$

The denominators of the Bayesian updating formula cancel out to obtain that

$$\frac{p}{1-p}\frac{\Pr(\text{4 out of 5} \mid \text{A succeeds})}{\Pr(\text{4 out of 5} \mid \text{A fails})} > \frac{p}{1-p}\frac{\Pr(\text{1 out of 1} \mid \text{B succeeds})}{\Pr(\text{1 out of 1} \mid \text{B fails})}.$$

Canceling and rewriting in terms of signal likelihoods given the iid assumption gives

$$\frac{{c_a}^4(1-c_a)}{{d_a}^4(1-d_a)} > \frac{c_b}{d_b}.$$

And by a similar calculation, if Alice chooses to pick B over A after collecting more information then it must be that the following holds

$$\Pr(\text{A succeeds} \mid \text{40 out of 50}) < \Pr(\text{B succeeds} \mid \text{10 out of 10}).$$

Which implies by an identical sequence of transformations that

$$\frac{{c_a}^{40}(1-c_a)^{10}}{{d_a}^{40}(1-d_d)^{10}} < \frac{{c_b}^{10}}{{d_b}^{10}}.$$

Note that the inequalities from the second decision are precisely that of the first taken to the power of 10. Therefore, if Alice's belief regarding the likelihoods, c_a, c_b, d_a, and d_b remained constant in the two decisions, her pattern cannot be rationalized as that of a Bayesian DM. While Bayesian updating may seem to be key, note we only required posteriors to be monotonic in the samples' likelihood ratios.

Before proceeding, two complementary features of the choice process should be highlighted for future sections. The first feature is that a sample with a small size is discounted potentially because it is perceived as noisy, and when the sample size increases, this concern disappears. This is precisely where confidence matters and where the iid assumption is violated. Under the iid assumption, with known likelihoods, the signal likelihoods are fixed and independent of the observed sample. And, therefore, leaves no room for their interpretation to change. We see instead that confidence increases in sample size. Further, the willingness to neglect the sample size occurs when one is confident in having observed a sufficiently large sample. The second feature is that, upon introspection and irrespective of the actual choices, one may realize that one was able to make these choices without knowledge of the signal likelihoods. Instead, one may have compared the sample characteristics. Taking this logic one step further, it suggests that one's choices may not be dependent on what one is told about signal likelihoods. I test the relevance of these features for decision-making experimentally and find evidence in favor of such a choice process. I also show theoretically that if the DM faces uncertainty regarding the likelihoods, then these findings are rationalized.

3 Theory

The DM chooses between two objects. Objects are assumed to be ex-ante identical and of binary quality, denoted by g and b for good and bad qualities respectively. For each object, the DM believes it has a probability p of being good. If the object chosen is of good quality, then she obtains a payoff of 1. If the object is bad, then she obtains instead a payoff of 0. For each object, the DM observes a sample of signals. Each signal can take on a finite set of types $t \in T$. A sample of signals is a T-dimensional vector with natural numbers as entries. Denote an object's sample by $s = (s^1, \dots, s^T) \in \mathbb{N}_0^T$, where s^t denotes the number of signals of type t in the object's sample. For example, each object could be a project, and a sample could be a set of predictions.

I study empirical content of models of updating and take the primitive of my framework to be a preference relation $\succeq$ defined on $\mathbb{N}_0^T \times \mathbb{N}_0^T$. Therefore, $s_1 \succeq s_2$ means to choose an object with sample s_1 over another object with sample s_2. This is taken to imply that the DM considers s_1 to be better evidence of an object being good than s_2. I now describe mental representations of a wide class of models of belief updating. I then offer their axiomatic characterization in terms of choice behavior between objects with samples.

In models of updating, the DM's belief regarding samples is based on her belief regarding individual signal realizations. Her belief regarding signals is described by a pair of likelihoods $\sigma_g = (\sigma_{g,1}, ..., \sigma_{g,T})$ and $\sigma_b = (\sigma_{b,1}, ..., \sigma_{b,T})$. Likelihoods σ_g and σ_b are her beliefs about the distribution over signal types conditional on the object being good and bad, respectively. For example, $\sigma_{g,t}$ denotes the probability a signal is of type t conditional on the object being good. I assume only that for all t $\sigma_{g,t} \in (0,1)$ and $\sigma_{b,t} \in (0,1)$, a full support condition.[8] I highlight that this is a very weak condition on beliefs as σs do not have to be correct, therefore allowing for model misspecification. Furthermore, I do not impose that these must add up to one, thus allowing for incoherent beliefs.

Given σ_g and σ_b, the DM can compute using the independence condition, for every sample, a likelihood ratio. For any sample s, its likelihood ratio is $L(s\ ;\ \sigma_g, \sigma_b) = \frac{\Pr(s \mid g)}{\Pr(s \mid b)} = \frac{\prod_{t=1}^T \sigma_{g,t}^{s^t}}{\prod_{t=1}^T \sigma_{b,t}^{s^t}}$. The DM uses an updating rule to update a posterior belief for each sample. I consider updating rules which are strictly monotonic in the likelihood ratio. While this seems like a strong assumption, it is satisfied by a wide range of non-Bayesian updating rules. See the online appendix for a more detailed discussion.[9]

Finally, once the DM has obtained a posterior belief for each object given its sample, she chooses the one with a higher posterior probability of being good. In this binary scenario, this amounts to any representation of choice under risk that satisfies FOSD. Therefore, non-EU theories such as rank-dependent EU or cumulative prospect theory are allowed. I note that often, a decision theorist wants to distinguish between different theories, which necessitates a large state space. My goal here is to investigate common behavioral implications of a general class of theories. Therefore, I look at the binary state space where these theories have identical predictions.

If a DM chooses according to the above, I say they have a likelihood ratio representation.

Definition 1. *A preference relation $\succeq$ has a likelihood ratio representation if there exist σ_g and σ_b such that for any samples s_1 and s_2,*

$$s_1 \succeq s_2 \text{ if and only if } L(s_1\ ;\ \sigma_g, \sigma_b) \geq L(s_2\ ;\ \sigma_g, \sigma_b).$$

[8]This condition is helpful for cosmetic purposes in the statement of the theorem and can be relaxed. It is neither substantive nor necessary for any results.

[9]To highlight why it is an intuitive assumption, consider the following scenario. The DM initially chose an object with sample s_1 over another with sample s_2. Then she learns that the signal-generating process is such that there is an additional signal type that she did not anticipate existed. This unanticipated signal type did not occur in either s_1 or s_2. She also learns that this additional signal type is equally likely for both good and bad objects. Therefore, this unanticipated and unobserved signal type is pure noise. Therefore, she should not change her choice. A violation of this assumption would imply that there are scenarios like the above where she would change her choice.

The likelihood ratio representation builds in two main assumptions: the updating rule being strictly increasing in the likelihood ratio, and that it is computed as if signals are iid with known likelihoods.[10] It turns out that such a representation has a simple axiomatization that is parallel to the EU representation of choice under risk. I introduce first the mixture operation and then my axioms.

Definition 2. *For* $\alpha \in [0,1]$, *samples* s_1 *and* s_2, *if* $\alpha s_1 \in \mathbb{N}_0^T$ *and* $(1-\alpha)s_2 \in \mathbb{N}_0^T$, *then the mixture* $s_1 \alpha s_2 = \alpha s_1 + (1-\alpha)s_2$.

Therefore $s_1 \alpha s_2$ denotes the sample that is obtained by adding α proportion of s_1 to $(1-\alpha)$ proportion of s_2. Because samples are vectors with natural numbers as entries, I restrict this definition to whenever both proportions are themselves samples. Given this definition, I define the axioms.

Axiom 1 (Separability). *For all* s_1 *and* s_2, *if* $s_1 \succeq s_2$ *then for any* s_3, $s_1 + s_3 \succeq s_2 + s_3$.

Axiom 2 (Mixture Independence). *For all* s_1 *and* s_2, *if* $s_1 \succeq s_2$ *then* $\forall \alpha \in (0,1)$ *and for any* s_3, $s_1 \alpha s_3 \succeq s_2 \alpha s_3$ *whenever* $\alpha s_1, \alpha s_2, (1-\alpha)s_3 \in \mathbb{N}_0^T$.

Axiom 3 (Continuity). *For all* s_1, s_2 *and* s_3, *the sets* $\{\alpha \mid \exists \kappa$ *such that* $\alpha\kappa s_1, (1-\alpha)\kappa s_2, \kappa s_3 \in \mathbb{N}_0^T$, *and* $(\kappa s_1)\alpha(\kappa s_2) \succeq \kappa s_3\}$ *and* $\{\alpha \mid \exists \kappa$ *such that* $\alpha\kappa s_1, (1-\alpha)\kappa s_2, \kappa s_3 \in \mathbb{N}_0^T$, *and* $(\kappa s_1)\alpha(\kappa s_2) \preceq \kappa s_3\}$ *are closed in* $\mathbb{Q} \cap [0,1]$.

Separability links a DM's preference over samples to the marginal effect of additional samples. In particular, it says that if an object with sample s_1 is chosen over another object with s_2, then for any sample s_3, the DM prefers an object with sample $s_1 + s_3$ to one with $s_2 + s_3$. Mixture independence is stated as under risk, with the caveat that the parts being mixed must be themselves samples as per the definition of the mixture operation. Finally, continuity is akin to the standard mixture continuity axiom under risk. The only differences are again due to the discreteness of the environment. First, as mixture proportion αs are rational numbers, the closure requirement is on the rationals as a subspace of $[0,1]$. Second, it is necessary to be able to multiply the sample sizes by arbitrarily large κ to find all the rationals that satisfy the condition.

I now present the representation theorem. Theorem 1 links the axioms to the likelihood ratio representation implied by models of updating and operationalizes it.

Theorem 1. *The following are equivalent:*

1. *The relation* $\succeq$ *has a likelihood ratio representation.*

2. *The relation* $\succeq$ *is transitive, complete, separable, and continuous.*

[10] I note also that it assumes the DM's mental model of the states is binary - which experimental researchers can assume safely but seldom verify. I show that relaxing this assumption while preserving the first two, yields behavioral patterns equivalent to relaxing signal likelihood being known. Therefore, as both relaxations have the same empirical content, it is up to the researcher to exercise his best judgment. In my experimental setting, it is safe to assume the mental state is binary given that subjects are explicitly told so. I show this in the Online Appendix.

3. *The relation $\succeq$ is transitive, complete, mixture independent, and continuous.*

4. *There exists a set $\{u_t\}_{t=1}^{T}$, and for all s_1 and s_2 we have*

$$s_1 \succeq s_2 \Leftrightarrow \sum_{t=1}^{T} u_t s_1^t \geq \sum_{t=1}^{T} u_t s_2^t.$$

The proof can be found in Appendix A.1. Theorem 1 links a broad class of models of updating with their empirical implications via the second statement. In particular, these models have choices satisfying separability. Therefore, the behavior exhibited in the thought experiment cannot be accommodated by updating rules that are strictly monotonic in the likelihood ratio, given that signals are perceived to be iid with known likelihoods. The third statement establishes the equivalence of separability and mixture independence. This gives a hint of the proof strategy. If the set of samples was on $\mathbb{R}_0^T$, then 3) $\Leftrightarrow$ 4) is immediate by the Mixture Space Theorem from Herstein and Milnor (1953). My proof proceeds by extending the domain of $\succeq$ to $\mathbb{Q}_0^T \times \mathbb{Q}_0^T$, allowing signal numbers to be rational numbers. This extension is carried out using the mixture operation. From there a generalization of the Mixture Space Theorem from Shepherdson (1980) can be applied.

Theorem 1 implies that a wide class of models of updating imply choice behaviors that have a linear utility representation. Therefore, these models predict indifference curves, drawn in the space of samples, must be parallel straight lines. If the signals have binary types, like in the thought experiment, then the space of samples can be illustrated in Figure 1. The x-axis and y-axis denote the number of bad and good signals, respectively. Therefore, any sample is a point on the plane. For graphical convenience, I showcase a choice pattern that is qualitatively identical but numerically different from the thought experiment. In Figure 1, $A1 = (1, 4)$, is initially chosen over $B1 = (0, 1)$. Then, the indifference curve through $B1$ must lie below $A1$ in the blue region but never on the dotted line if the preference is strict. Similarly, the indifference curve through $B2 = (0, 4)$, must lie above $A2 = (4, 16)$ in the red region if the DM switches when sample sizes are multiplied by 4. Note then the only way for indifference curves to be parallel is if they both lie on the dotted line. Therefore the thought experiment conflicts with any model with a likelihood ratio representation. In my experiment, I collect precisely such indifference curves and show that they are indeed not parallel straight lines but instead, rays that fan out as the thought experiment suggests.

I test models of updating characterized by Theorem 1 in the actual experiment. I note however that the thought experiment conflicts with a wider class of models. In particular, Theorem 1 holds for models satisfying four axioms. Whereas separability and transitivity are enough to conflict with the thought experiment. The theorem characterizes updating rules which are *strictly* monotonic in the likelihood ratio. However, weakly monotonic updating rules, such as Coarse Bayesian Updating (Jakobsen (2021)), are violated by the thought experiment even if they do not fall under Theorem 1. Table 13 summarizes known updating rules and their relationship with the

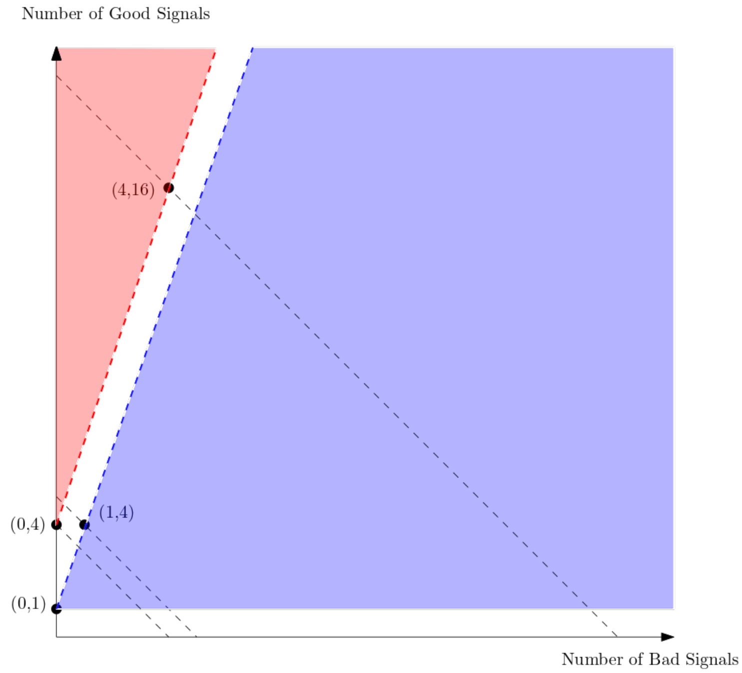


Figure 1: Indifference Curves Compatible with Thought Experiment

Table 1: Updating Rules and Relation with Actual and Thought Experiments

Updating Rules	Rejected By		Literature
	Actual Exp.	Thought Exp.	
Bayesian Updating	Yes	Yes	Bayes and Price (1763)
Grether Updating	Yes	Yes	Grether (1980); Möbius et al. (2022)
Weighted Bayesian	Yes	Yes	Epstein et al. (2010); Kovach (2021)
Divisible Updating	Yes	Yes	Cripps (2018)
Coarse Bayesian	No	Yes	Jakobsen (2021)
Confirmatory Bias	Yes	Yes	Rabin and Schrag (1999)
Size/Proportion Model	No	Yes	Griffin and Tversky (1992)
Inertial Updating	No	No	Dominiak et al. (2023)

actual experiment and the thought experiment.[11]

3.1 Confidence Elicitation

In this subsection, I introduce a confidence elicitation mechanism. The reader may skip to Section 4 where I show the experimental implementation of this mechanism. Additionally, my presentation here is restricted to confidence elicitation as it pertains to inference from samples. I generalize the framework and the mechanism for a wider class of choices in Appendix A.2.

The thought experiment hints that confidence in inference may be relevant for choice. However, according to earlier models of updating, the DM, given a sample, assigns an exact number to the posterior probability of an object being good. Then, when choosing between two objects, the DM

[11]The online appendix goes over each of these rules in detail.

knows with certainty which has a higher (subjective) probability of being good. These models therefore leave no room for the DM to have uncertainty about which object has a higher probability of being good. Therefore, to measure confidence and its relationship with choice, I first allow DMs to possess uncertainty regarding posteriors. This allows me to define lack of confidence as not knowing with certainty which object has a higher probability of being good. And I propose a confidence elicitation mechanism based on this definition.

Consider a DM who chooses between two objects with samples s_1 and s_2 respectively. To define confidence, I assume that the DM may be uncertain about the values of $\Pr(g|s_1)$ and $\Pr(g|s_2)$. I consider two common representations of this type of uncertainty. First, the DM could have a probability distribution P over values of $\Pr(g|s_1)$ and $\Pr(g|s_2)$. They then evaluate objects and their second-order distributions using some decision rule, examples include expected utility, smooth ambiguity (Klibanoff et al. (2005)), as well second-order forms of non-EU theories such as Segal (1990). Second, the DM may instead consider sets of probabilities Π_1, Π_2 as possible for posteriors $\Pr(g|s_1)$ and $\Pr(g|s_2)$. They then evaluate objects using a suitable decision rule such as maxmin EU (Gilboa and Schmeidler (1989)), or variational preferences (Maccheroni et al. (2006)). For both types of representations, I say the DM is *fully confident* in choosing an object with sample s_1 over one with sample s_2 if they believe with certainty that $\Pr(g|s_1) \geq \Pr(g|s_2)$. Formally, $P(\Pr(g|s_1) \geq \Pr(g|s_2)) = 1$ and $\min \Pi_1 \geq \max \Pi_2$ for the first and second class of models, respectively. Full confidence implies the DM assigns probability 1 to the object with sample s_1 having a higher probability of being good. Similarly, I say the DM *lacks confidence* when the above fails. My approach here is therefore general and incentive compatibility of my mechanism holds for a wide range of theories of confidence.

Suppose the DM is fully confident, then she has no instrumental value in learning whether $\Pr(g|s_1) \geq \Pr(g|s_2)$ is true as she already knows it. For any of the theories above, having full confidence implies zero value of information. Therefore, strictly positive willingness to pay to learn the correct action *only* occurs under lack of confidence. Using this channel, I consider the following elicitation mechanism:

- The subject is asked to choose between two objects with samples s_1, s_2 and a number $\delta \in [0, 1]$. Her payoff is determined as:
  1. With probability δ^2, they get a bad object.
  2. With probability $1 - \delta$, they get the object they chose.
  3. With probability $(1 - \delta)\delta$, they learn the object with the highest probability of being good, given s_1 and s_2, and can choose again.

Therefore, this mechanism gives the DM a chance to learn which object is statistically more likely to be good at a cost. Any theory of confidence, second-order probabilities, or sets of probabilities, assigns values V_2 and V_3 for the second and third options such that $V_2 \leq V_3$. Note a choice of δ yields a lottery over three outcomes: a bad object with value $V_1 < V_2$, an outcome with value V_2, and an outcome with value V_3. Choosing $\delta = 0$ yields a lottery with a guaranteed value of V_2. Therefore if the DM's evaluation of the risky option generated by the mechanism satisfies FOSD, then the DM chooses $\delta > 0$ only if $V_3 > V_2$. If one assumes expected utility over the uncertainty generated by δ and normalizes the value of a bad object to 0, one can solve for the optimal $\delta^* = \frac{1}{2}\frac{V_3 - V_2}{V_2}$. Therefore, choosing $\delta > 0$ implies a strictly positive instrumental value of information. Note that if the DM assigns $P(\Pr(g|s_1) \geq \Pr(g|s_2)) = 1$ or $\min \Pi_1 \geq \max \Pi_2$ then under any conventional updating rule for theories confidence, it must be that $V_2 = V_3$. Therefore, $\delta > 0$ implies the DM lacks confidence.

Proposition 1. *Suppose the DM's attitude regarding the lottery induced by the mechanism satisfies strict FOSD then $\delta > 0$ only if the DM lacks confidence.*

Proof in Appendix .

The presented mechanism requires the existence of an objectively correct choice that can be credibly signaled. However, one can get around this by providing a signal that the DM considers correlated with what they consider subjectively correct. For instance, in complex lottery choices, the expected value, and in dictator games, the average of other players' choices. If the non-instrumental value of information can be ruled out, then a DM chooses to acquire the signal only if they lack confidence and perceive the signal as informative.

The reader may also worry about the complexity of the lottery induced by the mechanism, as complexity has been shown to induce violations of FOSD. I note that this only makes the $\delta = 0$ case more attractive. Therefore this concern does not change the fact that $\delta > 0$ implies lack of confidence. This argument also holds for other theories where FOSD fails such as regret.

I show in the next section an implementation that is simple to understand for subjects and retains incentive compatibility. I also show in the experimental results section that the collected measure correlates well with an unincentivized measure.

4 Experimental Design

Overview. Subjects are told that there are 200 boxes, half of which are golden (good) and half are wooden (bad). Boxes also contain 10 colored balls in them. These balls are colored red or blue and the composition depends on the box's type. The relationship between color composition and box types differs across three between-subject treatments. Subjects are tasked with choosing between two boxes and go through three sets of choice tasks in random order. Subjects choose without knowing the boxes' types. But they may observe a sample of balls drawn with replacements from the boxes. After each choice, I elicit a measure of confidence. Subjects make, over the three sets of choice tasks, 16 choices in total.[12]

[12]Examples of the instructions, choice interface, and payment are provided in the online appendix.

Treatments. Subjects faced one of three treatments, which differed in the way the composition of balls in the box was determined. Two of the treatments have the color compositions fully determined by the box's type. In these treatments, as balls are drawn with replacements, they are iid conditional on the box's type. A third treatment involves uncertainty regarding the box's composition as the box's type does not fully determine it. Samples in this third treatment are therefore not iid conditional on the box's type. I summarize in Table 2 the compositions. Note for all the treatments, a red ball is a good signal while a blue ball is a bad signal.

Table 2: Summary of Treatment Types Given Box Type

Treatment Type	Golden Box	Wooden Box
Symmetric	7 red balls & 3 blue balls	7 blue balls & 3 red balls
Asymmetric	8 red balls & 2 blue balls	4 blue balls & 6 red balls
Correlated	4 red balls & 6 random balls	4 blue balls & 6 random balls

Note: "random" indicates balls equally likely to be red or blue, determined independently.

Choice Tasks. Each subject sees three sets of choices in random order. Two of the three sets of choices are called *comparative choice*. These involve choosing between boxes, for each of which the subject sees a sample of signals. These two sets differ in the number of total signals in each sample. A third set of choices is called *belief updating*. For this task, subjects choose between one box with a fixed chance of being golden and another with a sample of signals. From the two comparative choice tasks, I elicit 10 indifference curves. From the belief updating tasks, I elicit 4 indifference curves. All elicitations are done via a multiple-choice list where I elicit the subject's switching point. See Appendix B.3 for an example.

Comparative Choice Tasks. The choices from the two comparative choice tasks are as follows. Subjects are told that one box already drew a specific number of red balls out of 4. The other box has yet to draw any balls and subjects can choose based on the realized draw. For example, they can choose the first box whenever the second box draws less than 6 out of 10 red balls. The two sets of choice tasks differ in the number of balls drawn from this second box, which is either 10 or 25. I now elaborate on the specifics of the two tasks. Denote by (x, n) a box that drew x red balls out of n.

Size 4 vs Size 10: One set of ICs is elicited by asking for each $y \in \{0, 1, 2, 3, 4\}$ the number x_y of red balls such that $(x_y + 1, 10) \succeq (y, 4) \succeq (x_y, 10)$. Therefore $x_y + 1$ is the smallest number of red balls out of 10 that the subject deems to be better evidence of a golden box than y red balls out of 4. This gives me a bound for 5 indifference curves, and I use $x_y + 0.5$ as in the indifference point in my estimation whenever $x_y \neq 0$ or $x_y \neq 10$, in which case I use $x_y = 0$ and $x_y = 10$. In other words, I take $(x_y + 0.5, 10) \sim (y, 4)$ to hold whenever $x_y \notin \{0, 10\}$.

Size 4 vs Size 25: One set of ICs is elicited by asking for each $y \in \{0, 1, 2, 3, 4\}$ the number x_y

such that $(x_y + 1, 25) \succeq (y, 4) \succeq (x_y, 25)$. This gives me a bound for 5 indifference curves, one for each of $(y, 4)$. I use $x_y + 0.5$ as in the indifference point in my estimation whenever $x_y \neq 0$ or $x_y \neq 25$, in which case I use $x_y = 0$ and $x_y = 25$. In other words, I take $(x_y + 0.5, 10) \sim (y, 4)$ to hold whenever $x_y \notin \{0, 25\}$.

Recall that red balls are good signals in every treatment. Therefore monotonicity implies $(y, n) \succeq (y - 1, n)$, which implies $x_y \geq x_{y-1}$. I say a subject violates monotonicity if they display $x_{y-1} > x_y$ for any of the comparative tasks.

Belief Updating Task. In this task, subjects face one box with a fixed chance of being golden and another box that has yet to draw any signal. As in the comparative tasks, she can condition her choice on the realized draw. Denote by ℓ_y a box with y probability of being golden with $y \in \{0.25, 0.75\}$. I also elicit through 6 choice tasks x_ys such that $(x_y^4 + 1, 4) \succeq \ell_y \succeq (x_y^4, 4)$, $(x_y^{10} + 1, 10) \succeq \ell_y \succeq (x_y^{10}, 10)$ and $(x_y^{25} + 1, 25) \succeq \ell_y \succeq (x_y^{25}, 25)$. This gives me 4 indifference curves revealed through probabilistic equivalents. As before, I take the midpoint to be the point of indifference. This gives $(x_y^4 + 0.5, 4) \sim (x_y^{10} + 0.5, 10) \sim (x_y^{25} + 0.5, 25)$ whenever these midpoints are well defined, I use the extreme points of $0, 4, 10, 25$ if x_y ever equals these values.

Confidence Elicitation. Additionally, after each choice, the subject is given two options. I implement a simple form of my confidence elicitation mechanism. In particular, the subject is told that there is a statistically correct choice, which maximizes the probability of choosing a golden box. After each of the above choices, they are given two options:

(i) Always use the current choice.

(ii) 50% chance to learn the correct choice and can choose again, 49% chance to use the current choice, 1% chance of earning nothing.

Note that subjects are not guaranteed to learn the correct choice. Therefore, they are still incentivized, even if they choose option (ii), to give what they believe is the correct choice. Choosing option (ii) is a sufficient condition for the subject to perceive value in learning the correct choice. While it is not a necessary condition, it allows distinguishing between subjects who perceive a high enough value in learning the correct choice versus those who do not. I note finally that this learning occurs at the end of the experiment, therefore there is no risk of contamination from learning.

I also opt to inform the subjects of the statistically correct choice instead of replacing their choice. This is important as there may be subjects who wish to learn the statistically correct choice but not implement it. For instance, they may use it as a reference and then bias their own choice accordingly. This allows for a stronger test of lack of confidence.

I also collect, at the end of the study, an unincentivized, binary measure. Subjects are asked to report whether they believe they were close to the correct choice for most of the questions or not.[13]

Randomization and Order. Subjects are randomly assigned one of three treatments. Within the treatments, they are assigned a random order of blocks. The blocks are the two comparative choice tasks and the belief updating task. Within the blocks, to help with the consistency of choices, subjects always start by evaluating the box with the lowest value and each following box is the immediate next highest in value. Therefore, it is straightforward to respect monotonicity as a subject only needs to remember their last choice. Subjects are informed that one of their choices was randomly selected at the start of the study for payment. It is independent of their choices in the experiment. This theoretically eliminates hedging possibilities across tasks. Finally, the outcome of the confidence elicitation mechanism is only shown at the end of the experiment once the subject sees the task chosen for payment. This eliminates the mental burden of potentially having to learn and change their choices for many tasks, but more importantly, it prevents learning and contamination across questions.

5 Experimental Tests and Results

Background. I collected responses from 400 Prolific subjects. Subjects were paid \$2.5 USD for completing the study, with a chance to earn a bonus payment of \$5. The median completion time was 17 minutes, and around 60% of the subjects earned a bonus payment. Subjects were screened and had to pass a comprehension task. To participate, subjects needed an approval rate between 97%-99%, to have completed at least 100 studies, and to reside in the US. In the comprehension task, they are explicitly taught the monotonicity condition concerning red balls. Subjects can only start the actual tasks after demonstrating they understand the monotonicity condition. The study was pre-registered on Aspredicted.org.[14]

Variables and Measures. I focus on the indifference curves (ICs) and first study whether they are parallel straight lines in the aggregate and whether they differ by treatment. I then consider individual choices via three measures. The first measure quantifies whether an individual's ICs are parallel straight lines. For each indifference curve, I compute its angle relative to the x-axis. This yields 10 angles for each individual for the comparative choice tasks. The standard deviation of the angles should be close to 0 for straight parallel lines. Therefore, the larger this is, the less parallel the ICs must be. The second measure captures for each choice whether the subject chose according to the proportion of red balls (good signals) and neglected the sample size. For each choice, the subject chooses the minimal x_y^n, $n \in \{10, 25\}$, such that $(x_y^n, n) \succeq (y, 4)$, for each $y \in \{0, 1, 2, 3, 4\}$. I say that the subject's choice is consistent with a sample size neglect if $|\frac{x_y^n}{n} - \frac{y}{4}| \leq 0.05$. This implies

[13] See online appendix for an example.

[14] Please see here for pre-registration details.

Table 3: Summary Statistics

	Pooled		Symmetric		Asymmetric		Correlated	
	Full	Sub	Full	Sub	Full	Sub	Full	Sub
Standard Deviation of IC Angles	26.7	28.1	26.9	28.0	26.2	28.1	27.0	28.2
Sample Size Neglect (out of 10)	3.9	6.1	3.7	5.9	3.8	5.7	4.3	6.4
Opt to Learn (out of 10)	2.5	1.9	2.2	1.9	2.5	1.6	2.7	2.2
N	400	147	140	44	128	43	132	60

that the subject's choice is well predicted by the sample proportion. Figure 2 illustrates the type of ICs that would qualify. Note this is a demanding definition. For example, when $n = 10$ and $y = 2$, then the subject needs to pick exactly $x_y^n = 5$. Finally, for each choice, I collect a binary measure of confidence, as outlined in the previous section.

Analysis Summary. I first study the ICs elicited from the comparative choice tasks. I perform an individual analysis via the standard deviation of angles of ICs and reject separability for an overwhelming majority of subjects. Then, I investigate how these individual measures relate to each other. Building on the intuition from the thought experiment, I show first violations of separability can be accounted for via sample size neglect. Then I turn to my measure of confidence. I first corroborate its validity with the unincentivized measure. I show sample size neglect is positively associated with higher confidence. All the above are conducted using the comparative comparison tasks. I present at the end the ICs induced from the belief updating tasks and I show the same pattern emerges. As pre-registered, I will present results for the full sample as well as a sub-sample of subjects who did not violate the monotonicity condition. Non-violation is equivalent to having non-crossing ICs. In my data, 37% of subjects have 0 IC crossings, and they constitute this sub-sample. The theoretical maximum number of crossings is 8, and only 13% of subjects have more or equal to 4 crossings. I give some summary statistics of these variables in Table 3.

Table 3 show a few trends. On average, the subjects have high standard deviations for the angles of their ICs. On average, their ICs are not parallel straight lines. The average subject display choices consistent with sample size neglect 3.9 times out of 10. The sub-sample subjects display a much higher rate of sample size neglect, with 6.1 times out of 10 choices on average. I also find that the sub-sample is less likely to opt to learn and, therefore, more confident in their choices. Finally, treatment differences are not statistically significant except subjects are more likely to opt to learn in the correlated treatment compared to the symmetric treatment for the full sample.

Aggregate ICs. I plot in Figures 3, 4, and 5 the ICs of the three treatments for a Bayesian EU subject, the full sample, and the sub-sample, respectively. The aggregate ICs are not parallel for either the full or sub-samples. I can test whether the crossing points on the $N = 10$ and $N = 25$ lines are different between treatments. There are 3 treatments, with 10 such points, so this gives 30 tests. In the full sample, only 5 tests yield statistically significant differences between treatments at $p < 0.1$. For the sub-sample, only 6 tests yielded statistically significant differences. There are two takeaways. First, the aggregate ICs are not parallel straight lines. Therefore, suggesting that the models being tested do not account for aggregate behavior well. Second, subjects are essentially fully insensitive to the treatments This suggests that they are ignoring the likelihood and relying mostly on sample statistics such as the proportion of red balls and the total sample size.

Standard Deviation of IC angles. Figure 6 shows the distributions of standard deviations for the full sample and the sub-sample. In the full sample, only 5% and 20% of subjects have ICs with angles with a standard deviation below 10 and 20 degrees, respectively. In the sub-sample, only 6% and 17% of subjects have ICs with angles with a standard deviation below 10 and 20 degrees, respectively. Therefore, I conclude that the models are not only rejected at the aggregate level but also the individual level for almost all subjects. Using Kolmogorov-Smirnov tests, I investigate whether the distributions of standard deviations differ by treatment. I cannot reject the null for the full sample and sub-sample at any significance value $p \leq 0.10$. Finally, the spike at ≈ 33 is due to subjects who display sample size neglect for almost every choice.

Sample Size Neglect and non-Parallelism. As per my pre-registration, I explore the correlation between the standard deviation of angles of ICs and sample size neglect. The question I ask is: do people display non-parallel ICs because they are confused, choose noisily, or because they display sample size neglect, which is a systematic choice? The spike at 33 degrees in Figure 6 suggests the later channel and I give additional evidence here. I regress the standard deviation of angles, SD_i on the proportion of times, out of the 10 choices, a subject displays sample size neglect, P_i. Finally, X_i is a set of controls including sex, ethnicity, time taken (in the whole study), age as well as treatment dummies. I estimate regression (1) and the results are presented in Table 4. Regressions with treatment interaction terms can be found in Appendix B.2.

$$SD_i = \beta_0 + \beta_1 P_i + \lambda X_i + \epsilon_i \quad (1)$$

The results of the regression (Table 4) are in line with the aggregate ICs plotted earlier. There

Table 4: Non-Parallel ICs and Sample Size Neglect

	SD of Angles of ICs - SD_i			
	(1)	(2)	(3)	(4)
P_i	11.6 (0.94)	11.7 (1.00)	19.9 (1.8)	19.7 (1.9)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
R^2	0.19	0.20	0.47	0.48
N	400	379	147	140

Note: Robust standard errors in brackets.

is a strong correlation between sample size neglect and non-parallel ICs, which is stronger for the

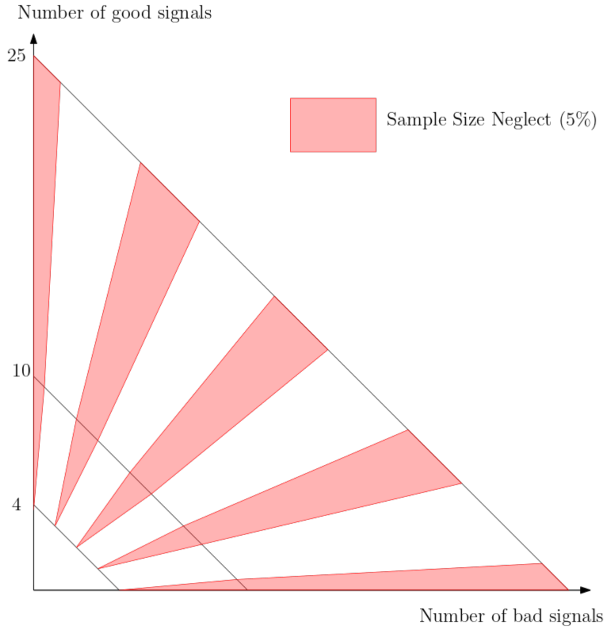


Figure 2: Sample Size Neglect ICs

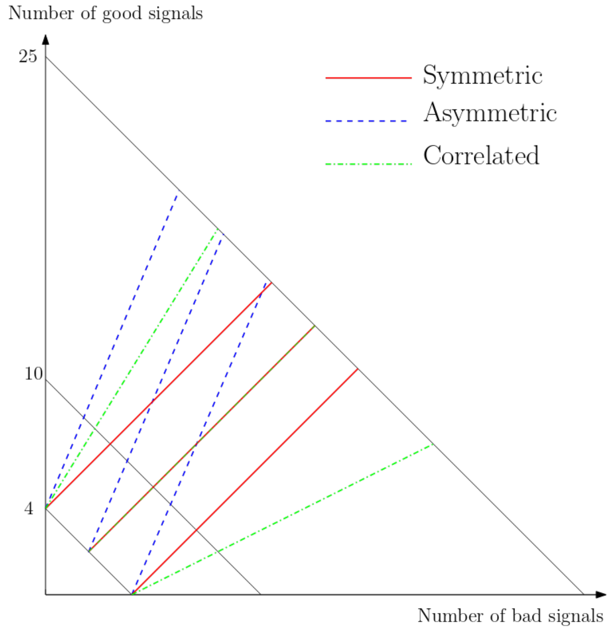


Figure 3: Bayesian EU IC

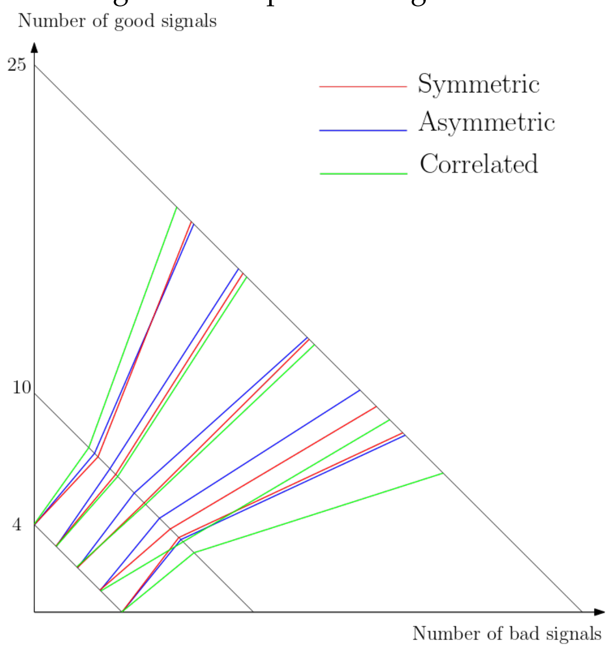


Figure 4: Full Sample IC

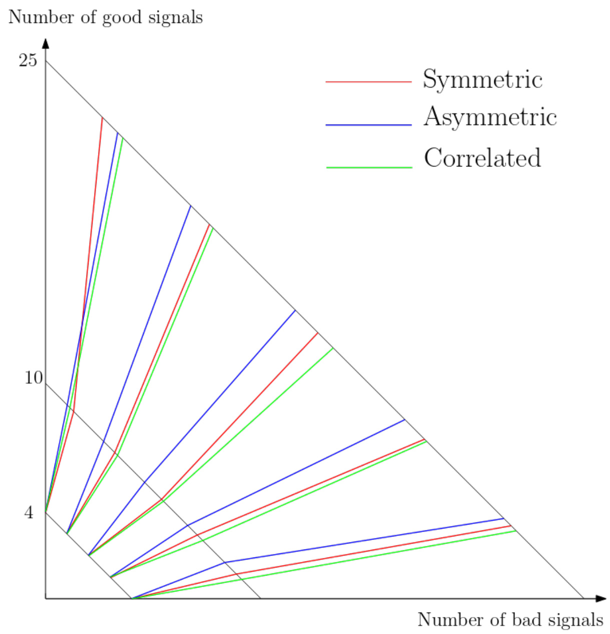


Figure 5: Sub-Sample IC

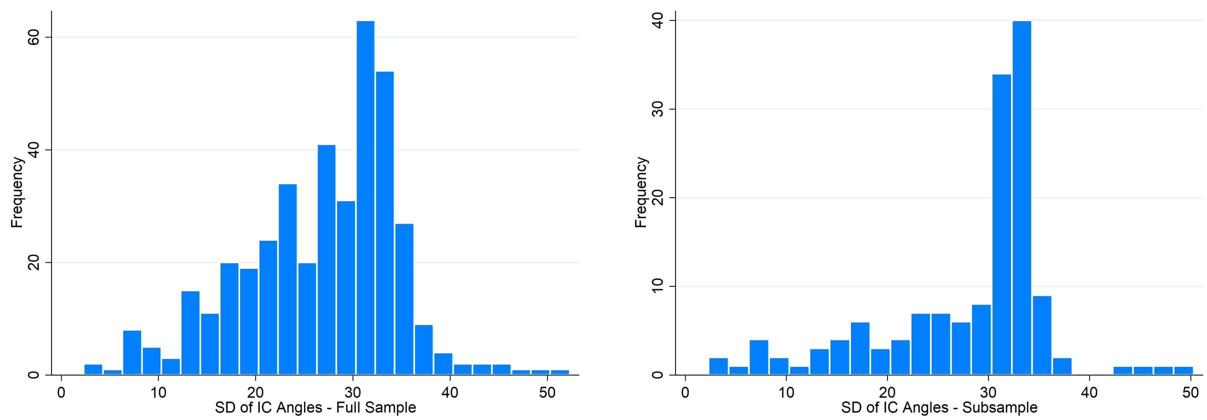


Figure 6: SD of IC Angles: Full Sample (left) and Sub-Sample (right)

sub-sample. On average, a DM who always displays sample size neglect has a standard deviation that is 12 and 20 degrees higher than a person who never displays sample size neglect for the full sample and sub-sample, respectively. These coefficients are statistically significant at $p < 0.01$ and do not differ when treatment dummies and controls are accounted for. Furthermore, the R^2s are high at 0.2 and 0.5 for the full sample and sub-sample, respectively. This does not rule out the possibility that subjects are choosing noisily. But allows me to conclude that a significant portion of non-parallelism and violation of models of updating is due to sample size neglect.

Confidence and Sample Size Neglect. I first perform a sanity check by verifying that the collected binary measure of confidence through my elicitation mechanism is highly correlated with the unincentivized self-reported confidence measure. Denote by O_i the percentage of times (out of 10) that a subject i opts to learn, so the higher this is, the less confident a subject is. And denote by C_i the binary self-reported measure. This self-reported measure is 1 if the subject reports believing in having chosen approximately correctly for most tasks, and is 0 otherwise. Finally, denote X_i a set of controls as well as treatment dummies. I run the following regression (2) as a linear probability model. The regression results are in Table 5. A subject who always opts to learn is, on average, 25% and 45% less likely to report that they are confident than someone who always opts out, in the full sample and sub-sample, respectively. Note that only 52% and 64% of subjects self-report to be confident in the full and sub-samples, respectively. Hence, the effects are significant both in magnitude and in statistical significance, as Table 5 shows. See Appendix B.3 for logit and probit results, which are consistent.

$$C_i = \beta_0 + \beta_1 O_i + \lambda X_i + \epsilon_i \quad (2)$$

Table 5: Self-Reported and Elicited Confidence

	Self-Reported Confidence - C_i			
	(1)	(2)	(3)	(4)
O_i	-0.26	-0.27	-0.47	-0.44
	(0.08)	(0.07)	(0.12)	(0.13)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
N	400	379	147	140

Note: Robust standard errors in brackets.

I test whether sample size neglect could be due to subjects not knowing how to choose and deferring their choices to the sample proportion. Figure 7 shows that whenever a choice displays sample size neglect, the subject is less likely to opt to learn for that choice. The effect is stronger for the larger samples (25) and for the sub-sample. For these choices, not displaying sample size neglect implies that the subject is 2.4 times more likely to opt to learn. For both the full and sub-sample this difference is statistically significant at $p < 0.05$, for all choice tasks in the 4 vs 25 rounds.[15] This suggests that sample size neglect is not due to confusion as subjects who display it are more confident in their choices. The differences are highly statistically significant for the large sample sizes.[16]

I also investigate this relationship in a linear regression. I denote choices by d and I set $o_d = 1$ if the subject opts to learn for that choice and $o_d = 0$ otherwise. Similarly, I set $p_d = 1$ if the choice d exhibits sample size neglect and $p_d = 0$ if it does not. Finally, X_i is a set of controls, including sex, ethnicity, time taken (in the whole study), age, and treatment dummies. To test whether sample neglect is related to lack of confidence, I consider the following specification (3). Table 6 presents the results for a linear probability model. Similar results are found for a logit and probit model. I also run the regression, as per my pre-registration, with interaction terms and found similar results.[17]

$$o_d = \beta_0 + \beta_1 p_d + \lambda X_i + \epsilon_d \quad (3)$$

Table 6: Sample Size Neglect and Confidence

	Opting to learn o_d			
	(1)	(2)	(3)	(4)
sample size neglect, p_d	-0.10	-0.10	-0.12	-0.11
	(0.013)	(0.014)	(0.022)	(0.026)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
N	4000	3860	1470	1410

Note: Robust standard errors in brackets.

[15] 7 out of 10 tasks has $p < 0.01$.

[16] Details are included in the online appendix.

[17] See Appendix B.3 for these additional regressions.

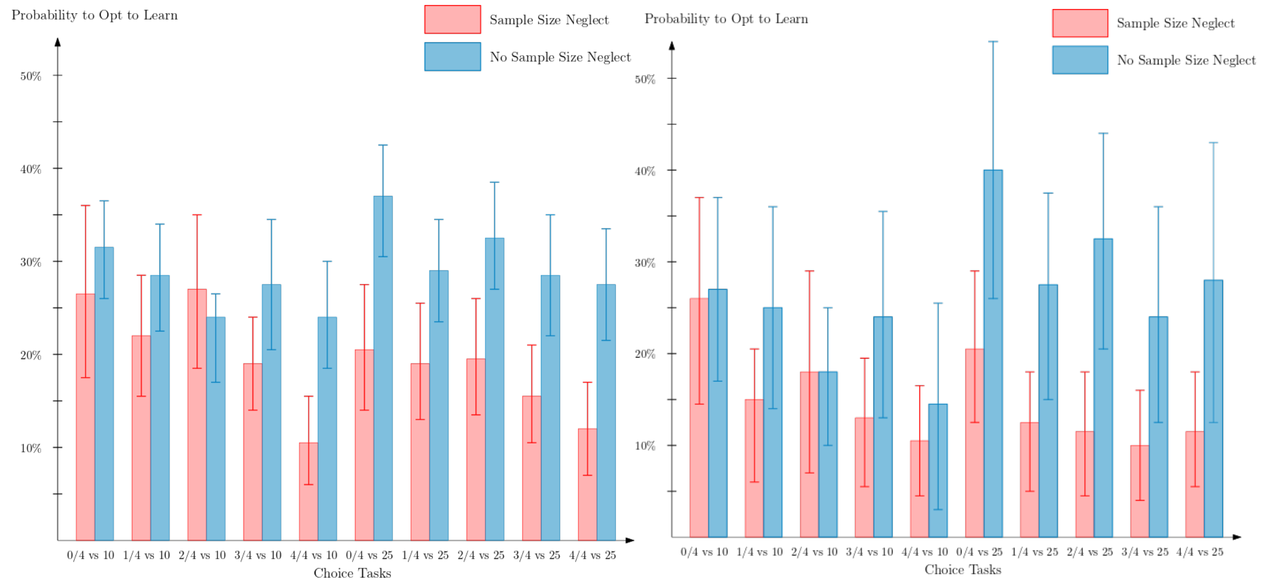


Figure 7: Full (Left) and Sub-Sample (Right) Confidence and Sample Size Neglect

In both the full sample and the sub-sample, if the subject's choice displays sample size neglect, then they are around 10% less likely to opt to learn. Note this is large as the average probabilities of opting to learn are 25% and 19% for the full and sub-samples, respectively. I conclude that sample size neglect is not due to a lack of confidence or noisy choices. On the contrary, a willingness to neglect the sample size and refer solely to sample proportions is associated with the subject being more confident. This is consistent with the intuition from the thought experiment. Section 6 provides a model that rationalizes this finding.

Belief Updating Tasks. The reader might wonder whether the comparativeness of the tasks pushes subjects to compare sample characteristics and ignore the signal likelihoods. To explore this possibility, I use the belief updating tasks to construct 4 indifference curves, presented below in Figure 8. The results are qualitatively similar, indifference curves still fan out, and further, the choices again do not vary by treatment in any significant manner.

Summary of Experimental Findings. Conventional models of updating are overwhelmingly rejected. Subjects are not sensitive to the information structure; I show that this is not driven by confusion. Rather, subjects intentionally choose by considering the sample characteristics. Many subjects display a sample size neglect bias. My confidence elicitation mechanism, which correlates well with an unincentivized measure, shows that sample size neglect bias is positively correlated with confidence.

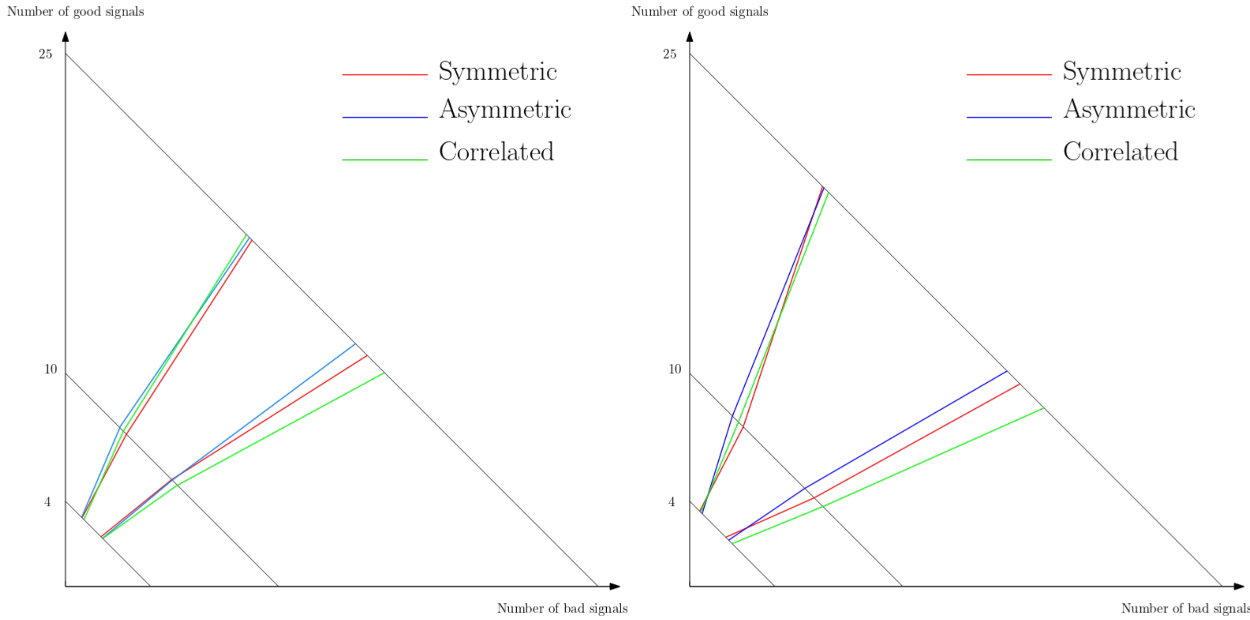


Figure 8: Full (Left) and Sub-Sample (Right) IC - Belief Updating

6 Likelihood Uncertainty and Signal Correlation

In this section, I discuss a natural class of belief about the signal-generating process that would accommodate the behavior of the thought experiment. Furthermore, I show that under this belief, it is asymptotically optimal to display sample size neglect.

Let us first reconsider the thought experiment. Recall the venture capitalist first elicited predictions from only one expert about project B, and this expert predicted success. Now, suppose they are to guess how likely an expert is to correctly predict success for project B. Mathematically, suppose they were asked to guess $\ell_1 = \Pr(\text{predicts success}|\text{B succeeds})$. Having only observed one signal, this is a difficult question to answer, and I doubt many readers would be willing to answer a high ℓ_1. However, suppose they now have observed 10 out of 10 experts predicting success. And recall they picked B over A, so they must believe that project B will succeed with a higher probability than A. Then, if asked again to guess $\ell_2 = \Pr(\text{predicts success}|\text{B succeeds})$, they must believe that the conditional term of "B succeeds" has a non-insignificant probability of being true, therefore, the empirical frequency observed, 10 out of 10, is at least somewhat indicative of the actual likelihood. This should incline a guess of $\ell_2 > \ell_1$. Note that the belief regarding the signal-generating process changes as one observes more signals. In other words, how one interprets signals is dependent on the sample one observes, so signals are thought to be correlated and not independent. In particular, there is some other uncertainty regarding the likelihood of signals, such as how hard it is to correctly predict success. These uncertainties are not fully known or determined by the underlying state, but as the sample size grows, the DM gradually learns about these and grows more confident.

Consider a simple binary state and binary signal type model. The state is good or bad, and signals are also good or bad. Therefore, samples are of the form $s_i = (s_{i,g}, s_{i,b})$, where the first entry denotes the number of good signals and the second entry denotes the number of bad signals. Let σ_g and σ_b denote the probability of a good signal conditional on the good and bad state, respectively. Similarly, $1 - \sigma_g$ and $1 - \sigma_b$ denote the probability of a bad signal conditional on a good and bad state, respectively. The DM assumes that σ_g and σ_b are drawn from CDFs F_g and F_b with convex support on $[0, 1]$. Therefore, the DM faces some uncertainty regarding the signal likelihood and believes the likelihoods to be distributed by F_g and F_b. Timing is important; the realization of σ_g and σ_b are determined first by F_g and F_b, and then the signals are drawn according to σ_g and σ_b. If different σ_g and σ_b are drawn for each signal, then there is no learning possible about this likelihood uncertainty, unlike as shown in the thought experiment. This case would then not be able to generate the behavior exhibited in the thought experiment for a Bayesian.

I illustrate first via a concrete example that relaxing the assumption that likelihoods are known accommodates the thought experiment.

Example 1. Suppose the venture capitalist does not know how good experts are at predicting different projects. This could be due to them not being an expert and unable to account for the difficulty of predicting accurately. Suppose the likelihoods for both predictions of A and B are randomly determined by $\sigma_g \sim F_g$ and $\sigma_b \sim F_b$ with $F_g = U[0.5, 1]$ and $F_b = U[0.3, 0.8]$. That is, they believe that if a project will succeed, then experts have at least a 50% chance of correctly predicting it. However, if a project cannot succeed, then they believe experts may be fooled, and potentially 80% could predict success. Then, the likelihood ratios of the two decisions display the switching patterns as desired.

Note that the sign switches precisely because they now have learned more about the likelihoods and are more confident, therefore, in what signals imply. Suppose the venture capitalist were to be asked the probability she believes each of these choices to be correct. Then, she would assign close to 1 to the second choice and strictly less to the first choice.

$$\frac{\Pr(\text{4 out of 5} \mid \text{A succeeds})}{\Pr(\text{4 out of 5} \mid \text{A fails})} = \frac{\int_{0.5}^{1} \sigma_g^4 (1-\sigma_g) d\sigma_g}{\int_{0.3}^{0.8} \sigma_b^4 (1-\sigma_b) d\sigma_b} > \frac{\int_{0.5}^{1} \sigma_g d\sigma_g}{\int_{0.3}^{0.8} \sigma_b d\sigma_b} = \frac{\Pr(\text{1 out of 1} \mid \text{B succeeds})}{\Pr(\text{1 out of 1} \mid \text{B fails})},$$

$$\frac{\Pr(\text{40 out of 50} \mid \text{A succeeds})}{\Pr(\text{40 out of 50} \mid \text{A fails})} = \frac{\int_{0.5}^{1} \sigma_g^{40} (1-\sigma_g)^{10} d\sigma_g}{\int_{0.3}^{0.8} \sigma_b^{40} (1-\sigma_b)^{10} d\sigma_b} < \frac{\int_{0.5}^{1} \sigma_g^{10} d\sigma_g}{\int_{0.3}^{0.8} \sigma_b^{10} d\sigma_b} = \frac{\Pr(\text{10 out of 10} \mid \text{B succeeds})}{\Pr(\text{10 out of 10} \mid \text{B fails})}.$$

For the rest of the discussion, I assume that $\succeq_{B,F}$ is the preference relation generated by a Bayesian EU DM who faces uncertainty $F = (F_g, F_b)$. I also assume that $\succeq_{B,F}$ additionally satisfies a weak monotonicity assumption. Monotonicity states that the DM recognizes good signals as good news and bad signals as bad news. I show that sample size neglect is asymptotically optimal irrespective of F given this assumption. Therefore, a DM who does not know what to believe about F_g and F_b but knows that monotonicity is satisfied by a Bayesian can do just as well as a Bayesian who knows F_g and F_b asymptotically by neglecting the sample size. In the following, I define first monotonicity and sample size neglect.

Definition 3. *A relation $\succeq$ is* monotonic *if* $\forall s_g, s_b \in \mathbb{N}_0, (s_g, s_b) \succeq_{B,F} (s_g - 1, s_b + 1)$.

Recall that objects are binary-valued, and utility can be normalized to 1 and 0. Denote by $\theta_1, \theta_2 \in \{g, b\}$ the object's types. Therefore, when given two objects with samples s_1, s_2, we have

$$U_{B,F}(s_1, s_2) = \max\left\{\Pr(\theta_1 = g \,|\, s_1, F), \Pr(\theta_2 = g \,|\, s_2, F)\right\},$$

which denotes the expected utility of the Bayesian EU DM who faces uncertainty F regarding likelihoods. We define the expected utility of a DM who uses the sample size neglect choice and faces F as follows:

$$U_{SSN}(s_1, s_2) = \begin{cases} \Pr(\theta_1 = g \mid s_1, F), \text{if } \frac{s_{1,g}}{s_{1,g}+s_{1,b}} > \frac{s_{2,g}}{s_{2,g}+s_{2,b}}, \\ \Pr(\theta_2 = g \mid s_2, F), \text{if } \frac{s_{1,g}}{s_{1,g}+s_{1,b}} < \frac{s_{2,g}}{s_{2,g}+s_{2,b}}, \\ \frac{1}{2}[\Pr(\theta_2 = g \mid s_2, F) + \Pr(\theta_1 = g \mid s_1, F)], \text{if } \frac{s_{1,g}}{s_{1,g}+s_{1,b}} = \frac{s_{2,g}}{s_{2,g}+s_{2,b}}. \end{cases}$$

Given that $U_{B,F}$ maximizes the choice's expected utility and U_{SNN} ignores the key statistical information provided from F and s_1, s_2, we have that $U_{B,F}(s_1, s_2) \geq U_{SSN}(s_1, s_2)$ in general. But the next result shows that asymptotically, the differences disappear.

Proposition 2. *If $\succeq_{B,F}$ is monotonic, then $\forall s_1, s_2 \in \mathbb{N}_0^2$, $\lim_{\kappa\to\infty} U_{B,F}(\kappa s_1, \kappa s_2) - U_{SSN}(\kappa s_1, \kappa s_2) = 0$.*

This proposition suggests why sample sizes are often ignored and why subjects can remain confident while ignoring sample sizes. Furthermore, it is consistent with our increasing comfort in ignoring the sample size and focusing on the proportion of good signals as sample sizes increase. The proof is contained in Appendix A.3, where I also show that the result is not restricted to binary signal types.

7 Conclusion

In this paper, I consider a DM who chooses between objects that are associated with samples. While this is a natural setting, I deviate from the literature on belief updating to study the empirical content of updating models in the context of samples. I theoretically characterize the empirical content of a wide class of models. Then, I illustrate a natural choice pattern which all these models fail to rationalize. These models are then tested and thoroughly rejected in a controlled experimental setting. The thought experiment suggests that the main discrepancy lies in that these models assume the DM is fully confident in how to interpret signals. Instead, subjects behave as if using a sample size neglect heuristic, which I show is asymptotically optimal whenever there is uncertainty regarding signal interpretation. Using a novel incentive-compatible confidence elicitation mechanism, I show that sample size neglect is positively correlated with confidence. This is predicted by a model of signal uncertainty and suggested intuitively by the thought experiment.

References

Barron, K. (2021). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics 24*, 31–58.

Bayes, M. and M. Price (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions (1683-1775) 53*, 370–418.

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

Caticha, A. and A. Giffin (2006). Updating probabilities. In *AIP conference proceedings*, Volume 872, pp. 31–42.

Chambers, C. P. and N. S. Lambert (2021). Dynamic belief elicitation. *Econometrica 89*(1), 375–414.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics 129*(4), 1625–1660.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics 22*(2), 369–395.

Cripps, M. W. (2018). Divisible updating. Working Paper.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2023). Inertial updating. Working Paper.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics 75*(4), 643–669.

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

Epstein, L. G. and Y. Halevy (2019). Ambiguous correlation. *Review of Economic Studies 86*(2), 668–693.

Epstein, L. G. and Y. Halevy (2023). Hard-to-interpret signals. *Journal of European Economics Association*. Accepted.

Epstein, L. G., J. Noor, and A. Sandroni (2010). Non-bayesian learning. *The BE Journal of Theoretical Economics 10*(1).

Epstein, L. G. and M. Schneider (2007). Learning under ambiguity. *Review of Economic Studies 74*(4), 1275–1303.

Frydman, C. and L. J. Jin (2022). Efficient coding and risky choice. *Quarterly Journal of Economics 137*(1), 161–213.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics 18*(2), 141–153.

Giustinelli, P., C. F. Manski, and F. Molinari (2022). Precise or imprecise probabilities? evidence from survey response related to late-onset dementia. *Journal of the European Economic Association 20*(1), 187–221.

Good, I. J. et al. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics 34*(3), 911–934.

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

Griffin, D. and A. Tversky (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology 24*(3), 411–435.

Halevy, Y., D. Walker-Jones, and L. Zrill (2023). Difficult decisions. Working Paper.

Herstein, I. N. and J. Milnor (1953). An axiomatic approach to measurable utility. *Econometrica 21*(2), 291–297.

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

Jakobsen, A. M. (2021). Coarse bayesian updating. Working Paper.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review 106*(4), 620.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica 77*(2), 603–606.

Karni, E. (2018). A mechanism for eliciting second-order beliefs and the inclination to choose. *American Economic Journal: Microeconomics 10*(2), 275–285.

Karni, E. (2020). A mechanism for the elicitation of second-order belief and subjective information structure. *Economic Theory 69*(1), 217–232.

Kellner, C., M. T. Le Quement, and G. Riener (2022). Reacting to ambiguous messages: An experimental analysis. *Games and Economic Behavior 136*, 360–378.

Kerwin, J. and D. Pandey (2023). Navigating ambiguity: imprecise probabilities and the updating of disease risk beliefs.

Khaw, M. W., Z. Li, and M. Woodford (2021). Cognitive imprecision and small-stakes risk aversion. *Review of Economic Studies 88*(4), 1979–2013.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005). A smooth model of decision making under ambiguity. *Econometrica 73*(6), 1849–1892.

Kovach, M. (2021). Conservative updating. Working Paper.

Liang, Y. (2023). Learning from unknown information sources. *Management Science*. Accepted.

Maccheroni, F., M. Marinacci, and A. Rustichini (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica 74*(6), 1447–1498.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2022). Managing self-confidence: Theory and experimental evidence. *Management Science 68*(11), 7793–7817.

Ngangoué, M. K. (2021). Learning under ambiguity: An experiment in gradual information processing. *Journal of Economic Theory 195*, 105282.

Nielsen, K. and L. Rigotti (2023). Revealed incomplete preferences. Working Paper.

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

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

Segal, U. (1990). Two-stage lotteries without the reduction axiom. *Econometrica*, 349–377.

Shepherdson, J. C. (1980). Utility theory based on rational probabilities. *Journal of Mathematical Economics 7*(1), 91–113.

Shishkin, D. and P. Ortoleva (2023). Ambiguous information and dilation: An experiment. *Journal of Economic Theory 208*, 105610.

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

Shore, J. and R. Johnson (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory 26*(1), 26–37.

Steiner, J. and C. Stewart (2016). Perceiving prospects properly. *American Economic Review 106*(7), 1601–1631.

Williams, P. M. (1980). Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science 31*(2), 131–144.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics 12*, 579–601.

Zhu, J., N. Chen, and E. P. Xing (2014). Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research 15*(1), 1799–1847.

A Proofs

A.1 Theorem 1

Throughout, I denote samples by x, y, z instead of s_1, s_2, s_3 to save a subscript. I denote by π_t^x the percent of signals of type t in x as well as N_x the total sample size of x.

$1) \Rightarrow 4)$. Pick any values of $\{\sigma_{g,t}, \sigma_{b,t}\}_{t \in T}$. Then note for any samples x and y, the DM always strictly prefers the one with a higher posterior. If the updating rule is strictly monotonic, which is equivalent to having a likelihood ratio representation, then the DM always strictly prefers a sample with a higher likelihood ratio. Then we see that

$$L(x) = \left[\prod_{t=1}^{T} \sigma_{g,t}^{\pi_t^x} / \prod_{t=1}^{T} \sigma_{b,t}^{\pi_t^x}\right]^{N_x} \text{ and } L(y) = \left[\prod_{t=1}^{T} \sigma_{g,t}^{\pi_t^y} / \prod_{t=1}^{T} \sigma_{b,t}^{\pi_t^y}\right]^{N_y}.$$

This then gives

$$L(x) > L(y) \text{ if and only if } N_x \sum_{t=1}^{T} \pi_t^x \log(\frac{\sigma_{g,t}}{\sigma_{b,t}}) > N_y \sum_{t=1}^{T} \pi_t^y \log(\frac{\sigma_{g,t}}{\sigma_{b,t}}).$$

Then choosing $u_t = \log(\frac{\sigma_{g,t}}{\sigma_{b,t}})$ shows that $1) \Rightarrow 4)$.

To show $4) \Rightarrow 1)$, suppose $\succeq$ is such that $\exists \{u_t\}$ such that

$$x \succeq y \text{ if and only if } N_x \sum_{t=1}^{T} \pi_t^x u_t \geq N_y \sum_{t=1}^{T} \pi_t^y u_t.$$

Then from above, we simply need to find λ_ts and $\alpha > 0$ such that the condition below holds. Then we can set $\sigma_{g,t} = \sigma_{b,t} \exp(\alpha u_t)$.

$$\forall t, \sigma_{b,t} \exp(\alpha u_t) \in (0,1) \text{ and } \sigma_{b,t} \in (0,1).$$

Note that this is simply a matter of scaling, as $\exp(\alpha u_t)$ is always positive. So we can always find a set of $\sigma_{b,t}$s small enough.

Now consider $2)/3)$ and $4)$, first $4) \Rightarrow 2)/3)$ is immediate by the functional form.

I start by showing that $3) \Rightarrow 4)$. Denote by $\mathbb{Q}$ the set of non-negative rationals. Then, $\mathbb{Q}^T$ is the set of samples with rational numbers of signals of each type. I define an extension of $\succeq$ on $\mathbb{Q}^T$, denoted by $\succeq^*$. Note that $\mathbb{Q}^T$ is what Shepherdson (1980) calls a *multiplier space* under mixtures with ratios in $\mathbb{Q}^T$. This is because, for any two rationals, their mixture by a ratio α that is itself a rational will be another rational. Shepherdson (1980) shows that $\succeq^*$ over such a space has a linear cardinal representation if and only if it satisfies three properties:

- Completeness+Transitivity.
- Closure of $\{\alpha \mid x\alpha y \succeq^* z\}$ and $\{\alpha \mid x\alpha y \preceq^* z\}$ in $\mathbb{Q}$ (as a subspace of $[0,1]$).

- Mixture: $x \succeq^* y$ implies $x\alpha z \succeq^* x\alpha y$.

Therefore, if we can extend $\succeq$ to $\succeq^*$ while giving it these properties, then we know there is a linear representation for $\succeq$.

Pick any $x, y \in \mathbb{Q}^T$, then they can be rewritten as $(\frac{x_1}{d}, ...\frac{x_T}{d})$ and $(\frac{y_1}{d}, ..., \frac{y_T}{d})$ where $\tilde{x} = (x_1, .., x_T) \in \mathbb{N}_0^T$ and $\tilde{y} = (y_1, .., y_T) \in \mathbb{N}_0^T$. I say $x \succeq^* y$ if and only if $\tilde{x} \succeq \tilde{y}$. Note that there is more than one way to rewrite it, but by mixture, these must agree under $\succeq$. Suppose $x = \tilde{x} \cdot \frac{1}{d} = \overline{x} \cdot \frac{1}{c}$ and $y = \tilde{y} \cdot \frac{1}{d} = \overline{y} \cdot \frac{1}{c}$. Let $d > c$, then if $\overline{x} \succeq \overline{y}$, by mixture, $\overline{x}\frac{c}{d}0 \succeq \overline{y}\frac{c}{d}0$, which is equivalent to $\tilde{x} \succeq \tilde{y}$.

Note $\succeq^*$ is complete by definition. For any two vectors, x and y, with rational numbers as entries, suffice to multiply them by $\Pi_{t=1}^T x_t y_t$ as the denominator to obtain $\tilde{x}, \tilde{y} \in \mathbb{N}_0^T$.

Consider transitivity of a triples, x, y, z samples. Then rewrite them as $x = \frac{\tilde{x}}{d}$, $y = \frac{\tilde{y}}{d}$, and $z = \frac{\tilde{z}}{d}$ where $\tilde{x}, \tilde{y}, \tilde{z} \in \mathbb{N}_0^T$. Then suppose $x \succeq^* y$, then $\tilde{x} \succeq \tilde{y}$ and similarly $\tilde{y} \succeq \tilde{z}$. So by transitivity of $\succeq$, $\tilde{x} \succeq \tilde{z}$ which implies $x \succeq^* z$.

Mixture is exactly like transitivity. Pick any x, y, α, z, we can rewrite all the terms as $\tilde{x}, \tilde{y}, \tilde{z}$. Then if $x \succeq^* y$, we have $\tilde{x} \succeq \tilde{y}$. Which implies $\tilde{x}\alpha\tilde{z} \succeq \tilde{y}\alpha\tilde{z}$, which implies $x\alpha z \succeq^* y\alpha z$.

Closure is directly given by the axiom. Note that $\{\alpha \mid x\alpha y \succeq^* z\}$ is the same as $\{\alpha \mid \forall\kappa, (\kappa x)\alpha(\kappa y) \succeq \kappa z\}$. This concludes that $\succeq$ has an extension $\succeq^*$, which has a linear utility form. Which implies $\succeq$ itself has such a representation. This concludes $3) \Rightarrow 4)$.

I now show $2) \Rightarrow 3)$ by showing separability and transitivity implies mixture. Suppose we have $x \succeq y$, then we want to show $x\alpha z \succeq y\alpha z$. Note firstly that separability implies that $x \succeq y$ and $x' \succeq y'$ then $x + x' \succeq y + y'$. This is done by three applications of separability plus transitivity of $\succeq$.

For mixture to be well defined, we have $\alpha x, \alpha y$ are samples of form $(\alpha x_1, ..., \alpha x_t)$ and $(\alpha y_1, ..., \alpha y_t)$ with integer entries. Note then that the smallest $\alpha = \frac{1}{N}$, which can work for both to be well defined, is when N is the largest common denominator of x_ts and y_ts. Similarly, any α that can work is of the form $\frac{k}{N}$. Note then suffice to show that $x \succeq y$ implies $\alpha x \succeq \alpha y$, then using separability with $(1 - \alpha)z$ yields mixture. First note that $\frac{1}{N}x \succeq \frac{1}{N}y$; if not, then we can apply separability on both sides and obtain $x \prec y$. Then this gives for any $\frac{k}{N}$ we have $\frac{k}{N}x \succeq \frac{k}{N}y$ as desired. This completes the proof.

A.2 Proposition 1

I consider a subject who must choose from a set of actions $a \in A$. The subject has a payoff function $\pi : A \times A \to Z$. The set Z denotes the potential consequences of her choices. It could be objective, e.g., monetary values, or subjective, e.g., subjective belief about the probability of winning, and I assume there is an outcome z_w that is understood by all subjects to be the worst one. $\pi(a, a^*)$ denotes the consequence should the subject choose a when a^* is the correct choice. Correct can be objective, as in the case of belief updating, or it could be subjective as in the case of dictator games or lottery choices. Finally, I denote by $s \in S$ a set of signal realizations. The subject may believe signals are correlated to the correct action a^*. I assume that $\pi(a, a^*)$ is uniquely maximized at $a = a^*$ for each a^*. This implies that not knowing the correct choice is payoff-relevant.

I note that these signals do not provide any value of information regarding uncertainties intrinsic to the experiment (such as lottery outcomes). The only instrumental value they can provide is in terms of the correctness of action ex-ante any resolution of uncertainty. The subject has some belief about the correct choice a^*. I say a subject is confident in knowing a^* whenever they assign probability 1 to some $a^* \in A$. If a subject is confident, then nothing can change her belief about a^*. Therefore, a confident subject should assign zero instrumental value to any signal, whether the subject believes it to be correlated with the correct action or not. The following example illustrates one common experimental setting that this framework nests.

Example 2. Consider eliciting a subject's probabilistic belief p that an event E occurred via some incentive-compatible mechanism, (Karni, 2009; Hossain and Okui, 2013). The correct belief, given the available information, is the Bayesian p^*. The subject reports p and is paid $\pi(p, p^*)$ that is uniquely maximized at $p = p^*$ whenever the elicitation is incentive-compatible. A set of signals could be to reveal to the subject the correct Bayesian posterior, in which case $S = [0, 1]$. Note this is only valuable if the subject is not confident that their report is the correct one.

Given the above setup, I propose the following confidence elicitation mechanism:

- The subject is asked to submit an action $a \in A$ and a number $\delta \in [0, 1]$.
  1. With probability δ^2, they get z_w.
  2. With probability $1 - \delta$, they get $\pi(a, a^*)$.
  3. With probability $(1 - \delta)\delta$, they observe a signal s and can change their action.

In the case of the correct choice being objective and known to the researcher, she can set $S = A$ and allow the signal to reveal the correct action. The procedure, in this case, allows the subject to be paid as if they knew the correct action a^* with some probability. The subject's belief about a^* may be a probability distribution over A or a set of possible a^*s depending on the theory of confidence that is chosen. Any such theory generates values V_2 and V_3 for the second and third outcomes of the above mechanism. Furthermore, any such theory can generate $V_2 < V_3$ only when the belief about a^* is not degenerate, and the signal is expected to be informative. I assume the DM's attitude towards the lottery generated by the mechanism satisfies FOSD. That is given any choice of δ, the DM faces a lottery with values $0, V_2$ and, V_3, normalizing the value of z_w to 0. Then FOSD implies that the DM picks $\delta > 0$ only if $V_2 > V_3$ which I show is only possible if she lacks confidence.

Proposition 3. *Suppose the DM's attitude regarding the lottery induced by the mechanism satisfies strict FOSD then $\delta > 0$ only if the DM lacks confidence.*

Proof: I will show that this holds for theories that assign either a probability over actions (over their correctness) and for theories that consider a set of actions to be correct. Suffice to show that $V_2 = V_3$ whenever the DM is confident in a^* being correct with probability 1. In this case, consider

S^* as the set of signals the DM expects to be possible given a^*. No conventional theory of updating can assign a positive probability to a state that previously had 0 probability.[18] Therefore, the value of observing any signal is 0, as the DM does not expect to change her belief. And if her attitude towards the mechanism is strictly FOSD then she does not acquire information.

A.3 Proposition 2

Proposition 2 is restricted to the binary signal case. I show here a more general T signal-type case. I consider a measure M of a sample, for instance, the average star rating or the percentage of good reviews, defined as follows.

Definition 4. *M is a* sample measure *if $M(s) = M(\kappa s)$ for $\kappa s \in \mathbb{N}_0^T$.*

A sample measure depends only on the distribution of signal types and not the sample size. When two samples have the same sample size, it is natural to use such a measure to choose. I show that if one chooses via such a measure when samples' sizes are equal, then one also chooses optimally when sizes are unequal but sufficiently large. Recall payoffs are normalized at 1 and 0 for good and bad objects. Denote by $U_B(s_1, s_2|F) = \max\{p(g|F, s_1), p(g|F, s_2)\}$ the utility of a Bayesian EU DM. Similarly denote by $U_M(s_1, s_2|F) = p(g|F, s_1)$ if $M(s_1) > M(s_2)$ and $U_M(s_1, s_2|F) = p(g|F, s_2)$ if $M(s_1) < M(s_2)$, in case $M(s_1) = M(s_2)$, $U_M(s_1, s_2|F) = \frac{1}{2}[p(g|F, s_1) + p(g|F, s_2)]$. U_M is the expected utility of a DM who uses a M measure heuristic for her choices.

Proposition 4. *Let F and M be such that $\forall |s_1| = |s_2|,\ p(g|F, s_1) > p(g|F, s_2)$ if and only if $M(s_1) > M(s_2)$, then $\lim\limits_{\kappa\to\infty} U_B(\kappa s_1, \kappa s_2|F) - U_M(\kappa s_1, \kappa s_2|F) = 0$.*

Note if M is taken to be the sample proportion of success, then the precondition $\forall |s_1| = |s_2|,\ p(g|F, s_1) > p(g|F, s_2)$ is implied by monotonicity. Therefore, this proposition implies Proposition 2.

Proof: Note first that any sample s gives an empirical likelihood $\sigma^s = [\frac{s^1}{|s|}, .., \frac{s^T}{|s|}] = [\sigma_1^s, ..., \sigma_T^s]$, I first show the following lemma. Denote by f_g, f_b the pdfs of F.

Pick any signal x and denote by Σ the set of T-dimensional likelihoods then the following

[18]The only exceptions are theories that consider observing signals which are not possible given the DM current belief such as Ortoleva (2012). However, the DM does not expect to receive such signals and hence in her information acquisition decision she does not account for these.

holds.

$$\begin{aligned}
\lim_{N\to\infty}\frac{p(g|x,F)}{p(b|x,F)} &= \lim_{N\to\infty}\frac{p(g)}{p(b)}\frac{\int_\Sigma f_g(\boldsymbol{\sigma})[\sigma_1^{\sigma_1^x}...\sigma_T^{\sigma_T^x}]^N d^T\boldsymbol{\sigma}}{\int_\Sigma f_b(\boldsymbol{\sigma})[\sigma_1^{\sigma_1^x}...\sigma_T^{\sigma_T^x}]^N d^T\boldsymbol{\sigma}},\\
&= \lim_{N\to\infty}\frac{p(g)}{p(b)}\frac{\int_\Sigma f_g(\boldsymbol{\sigma})e^{N[\sum_{i=1}^T\sigma_i^x\ln(\sigma_i)]}d^T\boldsymbol{\sigma}}{\int_\Sigma f_g(\boldsymbol{\sigma})e^{N[\sum_{i=1}^T\sigma_i^x\ln(\sigma_i)]}d^T\boldsymbol{\sigma}},\\
&= \lim_{N\to\infty}\frac{p(g)}{p(b)}\frac{(\frac{2\pi}{N})^{\frac{T}{2}}\frac{f_g(\sigma^{g,x})e^{N\sum_{i=1}^T\sigma_i^x\ln(\sigma_i^{g,x})}}{|-H(f_g)(\sigma^{g,x})|^{\frac{1}{2}}}}{(\frac{2\pi}{N})^{\frac{T}{2}}\frac{f_b(\sigma^{b,x})e^{N\sum_{i=1}^T\sigma_i^x\ln(\sigma_i^{b,x})}}{|-H(f_b)(\sigma^{b,x})|^{\frac{1}{2}}}},\\
&= \begin{cases}\frac{p(g)}{p(b)}\frac{f_g(\sigma^{g,x})}{f_b(\sigma^{b,x})} & \text{if } \sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{b,x})=\sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{g,x}),\\ \infty & \text{if } \sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{b,x})<\sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{g,x}),\\ 0 & \text{if } \sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{b,x})>\sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{g,x}).\end{cases}
\end{aligned}$$

Where $\sigma^{g,x} = \arg\max_{f_g(\sigma)>0}\sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{g,x})$ and $\sigma^{b,x} = \arg\max_{f_b(\sigma)>0}\sum_{t=1}^T\sigma_t^x\ln(\sigma_t^{b,x})$. Line 2 to 3 is by Laplace's method. $H(f_g)(\sigma^{g,x})$ is the determinant of the Hessian of f_g evaluated at $\sigma^{g,x}$ so it is finite. Laplace's method requires a unique maximizer, which may not occur. I show that this can be circumvented. Note that if x has strictly positive observations for each signal type, then $\sum_{i=t}^T\sigma_t^x\ln(\sigma_t^{g,x})$ is strictly quasi-concave in $\sigma_t^{g,x}$, suffice to note $\ln(\alpha z+(1-\alpha)w)>\alpha\ln(z)+(1-\alpha)\ln(w)$ by strict concavity. Now suppose x has signals types with zero observations, then note that any two accuracies σ^1,σ^2 which assign the same values to the non-zero types will satisfy $\sum_{t=1}^T\sigma_i^x\ln(\sigma_t^1)=\sum_{t=1}^T\sigma_t^x\ln(\sigma_t^2)$. So, we can "compress" the signal type space and rewrite F_g,F_b so that the maximizer is unique. Denote by T^x the set of types for which x has zero observation. Then, define the signal space Σ^* to have for types where x has non-zero observation and a type t^*. Then, defining F_g,F_b accordingly will yield distributions where the function has a unique maximizer.

Take any x and y, then note that as sample size grows, the two ways for $U_B(\kappa x,\kappa y|F)-U_M(\kappa x,\kappa y|F)>0$ to hold is if 1) both x,y are in the first category and the heuristic M orders them incorrectly or 2) x,y are in different categories respectively and the heuristic M orders them incorrectly. If we have both x,y in category 2 or 3, then asymptotically picking either has the same payoff, so $U_B=U_M$.

Then take any x,y and suppose x,y are both in the first category. Suppose wlog that $M(x)>M(y)$. Consider $w=\kappa|x|y$ and $z=\kappa|y|x$; note these two have the same sample size, and we can make these arbitrarily big. Then we have $w\succ z$ which gives $\frac{f_g(\sigma^{g,x})}{f_b(\sigma^{b,x})}>\frac{f_g(\sigma^{g,y})}{f_b(\sigma^{b,y})}$. So, the Bayesian choice coincides with the heuristic choice. Take any x,y, in two different categories; suppose wlog that the Bayesian posterior of x converges to 1 while that of y converges to 0. Then note $M(x)>M(y)$ by the same argument as above. The same argument applies to other cases.

B Additional Regressions and Multiple Price List Example

B.1 Regression with Interaction Terms

I regress specifications (1) and (2) with treatment interaction as

$$SD_i = \beta_0 + \sum_{t \in T} \delta_t \beta_{1,t} P_i + \sum_{t \in T} \delta_t D_t + \lambda X_i + \epsilon_i, \text{ and}$$
$$o_d = \beta_0 + \sum_{t \in T} \delta_t \beta_{1,t} p_d + \sum_{t \in T} \delta_t D_t + \lambda X_i + \epsilon_d.$$

Table 7: Non-Parallel ICs and Sample Size Neglect

	SD of Angles of ICs - SD_i			
	(1)	(2)	(3)	(4)
$\beta_{1,sym}$	12.5	12.0	19.9	21.0
	(1.2)	(1.8)	(1.8)	(2.5)
$\beta_{1,asy}$	10.7	11.3	19.9	17.4
	(1.2)	(1.7)	(1.8)	(3.9)
$\beta_{1,cor}$	11.6	12.3	19.9	20.9
	(1.0)	(1.6)	(1.8)	(3.8)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
R^2	0.19	0.22	0.47	0.49
N	400	386	147	141

Note: Robust standard errors in brackets.

Table 8: Opting to Learn and Sample Size Neglect

	Opt To Learn - o_d			
	(1)	(2)	(3)	(4)
$\beta_{1,sym}$	-0.13	-0.10	-0.12	-0.08
	(0.02)	(0.02)	(0.03)	(0.04)
$\beta_{1,asy}$	-0.09	-0.08	-0.16	-0.11
	(0.02)	(0.02)	(0.03)	(0.04)
$\beta_{1,cor}$	-0.09	-0.12	-0.10	-0.13
	(0.2)	(0.02)	(0.03)	(0.04)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub

Note: Robust standard errors in brackets.

B.2 Logit and Probit Regressions for Specification (2) and (3)

Table 9: Self-Report and Elicited Confidence - Logit

	Self-Reported Confidence - C_i			
	(1)	(2)	(3)	(4)
O_i	-1.11	-1.26	-1.91	-1.95
	(0.34)	(0.35)	(0.60)	(0.62)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
N	400	391	147	146

Note: Robust standard errors in brackets.

Table 10: Self-Report and Elicited Confidence - Probit

	Self-Reported Confidence - C_i			
	(1)	(2)	(3)	(4)
O_i	-0.69	-0.79	-1.18	-1.20
	(0.21)	(0.21)	(0.37)	(0.38)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
N	400	386	147	141

Note: Robust standard errors in brackets.

Table 11: Sample Size Neglect and Confidence- Logit

	Opting to learn o_d			
	(1)	(2)	(3)	(4)
p_d, sample size neglect	-0.53	-0.57	-0.73	-0.71
	(0.08)	(0.14)	(0.08)	(0.14)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
N	4000	3860	1470	1410

Note: Robust standard errors in brackets.

Table 12: Sample Size Neglect and Confidence - Probit

	Opting to learn o_d			
	(1)	(2)	(3)	(4)
p_d, sample size neglect	-0.31	-0.33	-0.41	-0.41
	(0.05)	(0.05)	(0.08)	(0.08)
Controls/Treat.Dummy	No	Yes	No	Yes
Full/Sub-Sample	Full	Full	Sub	Sub
N	4000	3860	1470	1410

Note: Robust standard errors in brackets.

B.3 Multiple Price List Example

Choice

Hover to see the experimental set-up.

You are offered a choice between two boxes, A and B. Each of these boxes were randomly picked from the 200 boxes. You will be paid $5 if the box you pick is golden and the computer has selected this task for payment. Box A and B were randomly picked the following way:

1. Box A was picked as follows:
   1. The computer drew 4 balls from each of the 200 boxes. They were drawn one at a time, returning each ball to the box after it was drawn.
   2. 5 out 200 boxes had 0 out 4 red balls drawn from them (4 other balls were blue).
   3. Box A was randomly selected from these 5 boxes.

2. Box B was picked as follows:
   1. The computer randomly picked Box B from the 200 boxes.
   2. The computer will draw 10 balls from Box B, one at at time and returning the drawn ball to the box.
   3. You can make your choice based on the number of red balls that are drawn from B.

Choose Box A			Choose Box B
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **0 out of 10 (0%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **1 out of 10 (10%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **2 out of 10 (20%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **3 out of 10 (30%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **4 out of 10 (40%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **5 out of 10 (50%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **6 out of 10 (60%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **7 out of 10 (70%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **8 out of 10 (80%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **9 out of 10 (90%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○	○	Box B - if **10 out of 10 (100%)** balls drawn are red.

You may wonder if there is a "correct" choice to this task. Using statistical theory, the computer can calculate when Box A or B is more likely to be golden. You may not know what the correct choice is, therefore, you are offered after each choice a chance to learn the correct choice and change your choices. In particular, you can choose between the following:

Choose an option:

Next

Figure 9: Example of MPL Choice

A Online Appendix

A.1 Some Further Implementation Subleties

Before moving on, I discuss three subtleties about the implementation of the mechanism.

First, when there is an objectively correct action, one may wonder if it is better to offer subjects a chance to replace their action with the objectively correct one. The answer is no because subjects may not perceive the objectively correct answer as payoff maximizing. However, they may believe (erroneously) that the objectively correct action is related to the subjectively correct action, in which case there is still gain in learning it and less gain in the action being replaced. For instance, consider a subject who learns that the Bayesian posterior is 0.99. She may consider that to be too extreme and report 0.7. For such a subject, she may still find value in learning the Bayesian posterior but be unwilling to replace her report with the Bayesian posterior.

Second, the cost they incur is a probability of obtaining the z_w outcome. The cost is probabilistic to guarantee incentive compatibility for non-risk-neutral individuals. For risk-neutral individuals, imposing a flat fee can be optimal.

Third, the signal can be used to elicit the source of lack of confidence. For instance, consider a subject who is not confident in her choice between lotteries. Some theories explain this as the subject having difficulty in computing the expected value, while other theories highlight the subject's uncertainty regarding her own risk attitudes. To test the first theory, the signal offered could be simply the expected value. If subjects are willing to pay for it, then it must be that the signal is valuable in clarifying uncertainty regarding a^*. Similarly, if a subject is uncertain of her own risk attitude, perhaps they will know better after making other choices. Option 3 could be simply the possibility of coming back to this choice.

B Online Appendix: Theoretical Notes

B.1 Updating Rules Details

I first begin with a proposition that shows that the thought experiment conflicts with updating rules that are weakly monotone in the likelihood ratio. Then, I individually analyze the above rules.

Proposition 5. *A preference relation $\succeq$ is said to display "switching" if $\exists s_1, s_2, \kappa$ such that $s_1 \succ s_2$ and $\kappa s_1 \prec \kappa s_2$. A preference relation $\succeq$ is said to be derived from an updating rule that is weakly monotonic in the likelihood ratio if $\exists \sigma_g, \sigma_b$ such that $L(s_1|\sigma_g, \sigma_b) \geq L(s_2|\sigma_g, \sigma_b)$ implies $s_1 \succeq s_2$. If $\succeq$ is derived from an updating rule that is weakly monotonic in the likelihood ratio, then it cannot display switching.*

Proof: Suppose $\succeq$ is derived from a rule that is weakly monotone in the likelihood ratio. Then $s_1 \succ s_2$ implies $L(s_1|\sigma_g, \sigma_b) > L(s_2|\sigma_g, \sigma_b)$. Then, if it displays switching, we must have some $s_1 \succ s_2$ and yet $\kappa s_2 \succ \kappa s_2$. So we must have $L(s_1|\sigma_g, \sigma_b)^\kappa < L(s_1|\sigma_g, \sigma_b)^\kappa$ which contradicts the earlier statement.

Table 13: Updating Rules and Relation with Actual and Thought Experiments

Updating Rules	Rejected By		Literature
	Actual Exp.	Thought Exp.	
Bayesian Updating	Yes	Yes	Bayes and Price (1763)
Grether Updating	Yes	Yes	Grether (1980); Möbius et al. (2022)
Weighted Bayesian	Yes	Yes	Epstein et al. (2010); Kovach (2021)
Divisible Updating	Yes	Yes	Cripps (2018)
Coarse Bayesian	No	Yes	Jakobsen (2021)
Confirmatory Bias	Yes	Yes	Rabin and Schrag (1999)
Size/Proportion Model	No	Yes	Griffin and Tversky (1992)
Inertial Updating	No	No	Dominiak et al. (2023)

I now provide some additional details for the updating rules of Table 13. I show that a wide class of models of non-Bayesian updating are functions of the likelihood ratio in this binary state setting. Denote by $\ell_g(s) = p(s|g, \sigma)$ where s is a sample. And denote by $p_B(g|s)$ the Bayesian posterior.

Bayesian updating: The Bayesian posterior ratio is proportional to the likelihood ratio and also, therefore, strictly increasing.

Grether updating: $p(g|s) = \frac{p(g)^\beta \ell_g(s)^\delta}{p(g)^\beta \ell_g(s)^\delta + (1-p(g))^\beta \ell_b(s)^\delta}$.

Note the posterior ratio of the states is:$[\frac{p(g)}{1-p(g)}]^\beta [\frac{\ell_g(s)}{\ell_b(s)}]^\delta$ where $\delta \geq 0$ is the signal reaction term. Therefore, the posterior ratio is weakly increasing and a function of the likelihood ratio whenever $\delta \geq 0$. If $\delta > 0$, which is the standard estimate, otherwise the DM is ignoring information, then it is strictly increasing.

Motivated Beliefs: $p(g|s) = \alpha p^* + (1-\alpha) p_B(g|s)$.

This updating rule is a convex combination of the Bayesian posterior and some arbitrary belief p^*. Taking p^* to be the prior would lead to underreaction. As the Bayesian posterior increases in the likelihood ratio, this updating rule is also. Similarly, whenever $\alpha < 1$, whenever the DM does update, it is strictly increasing.

Divisible Updating: It is shown in Cripps (2018) that a divisible updating rule must be homogeneous of degree 0 to the likelihoods of a signal. Therefore, if two samples have the same likelihood ratio, they will have the same posterior under a divisible updating rule. So, the updating rule is a function of the likelihood ratio.

Coarse Bayesian: this updating rule stipulates that there are convex subsets $P_1, .., P_N$ of $[0, 1]$ each with a "representative" probability $p_1 \in P_1, ..., p_n \in P_n$. The updating rule says that if $p_B(g|s) \in P_i$, then $p(g|s) = p_i$. So if the Bayesian posterior is in P_i, then the posterior is p_i. As convex subsets must be intervals, this updating rule is a function of the Bayesian posterior, which is a function of the likelihood ratio. This updating rule is not strictly increasing but weakly increasing and, therefore, cannot account for the thought experiment.

Size/Proportion Model: Griffin & Tversky propose, for a restrictive environment a regression that attempts to capture the weight of proportion of good signals, π, and sample size, N, in a DM's belief updating. Their regression only applies with prior $p = 0.5$ and when the signal structure is symmetric with $\sigma_g = 1 - \sigma_b$. So technically this is not an updating rule. Their regression can be mapped as an updating rule in this specific setting. In particular, they estimate:

$$\ln(\ln(\frac{p(g|s)}{1 - p(g|s)})) = \alpha_1 \ln(2\pi - 1) + \alpha_2 \ln(N) + \epsilon.$$

The idea here is that $\alpha_1 = \alpha_2$ implies Bayesian updating when the prior is uninformative $p(g) = 0.5$. Note that this is not separable but can still not accommodate the thought experiment. If two samples of size N_1, N_2 are multiplied to κN_1, κN_2, then they have a $+\alpha_2 \ln(\kappa)$, and therefore any inequalities are preserved.

Confirmatory Bias: This is technically a special type of perception rule; my setting is a little different as signals arrive together in one batch, whereas they model sequential observation with binary signals. However, the sequence turns out not to matter, so a faithful way of importing their model is to assume the DM has a bias for a state and may misperceive a signal for the other state as a signal for the biased state with probability q. Therefore, each sample (π, N) is changed to $(\pi(1-q), N)$ or $(\pi + (1-\pi)q, N)$, the rest is Bayesian updating.

Suppose a DM uses such an updating rule and perceives signals as iid. Then, by Theorem 1, the updating rule is strictly monotonic if and only if the $\succeq$ is separable. As we already have transitivity, completeness, and continuity.

Let $x = (\pi_1, N_1)$ and $y = (\pi_2, N_2)$. If $x \succeq y$ then the DM's belief σ_g, σ_b and bias in updating $q \in (0, 1)$ satisfy

$$\frac{\sigma_g^{\pi_1(1-q)N_1}(1-\sigma_g)^{(1+\pi_1(q-1))N_1}}{\sigma_b^{\pi_1(1-q)N_1}(1-\sigma_b)^{(1+\pi_1(q-1))N_1}} \geq \frac{\sigma_g^{\pi_2(1-q)N_2}(1-\sigma_g)^{(1+\pi_1(q-1))N_2}}{\sigma_b^{\pi_2(1-q)N_2}(1-\sigma_b)^{(1+\pi_2(q-1))N_2}}.$$

This then implies that

$$\pi_1(1-q)N_1 \ln(\frac{\sigma_g}{\sigma_b}) + (1+\pi_1(q-1))N_1 \ln(\frac{1-\sigma_g}{1-\sigma_b}) \geq \pi_2(1-q)N_2 \ln(\frac{\sigma_g}{\sigma_b}) + (1+\pi_2(q-1))N_2 \ln(\frac{1-\sigma_g}{1-\sigma_b}).$$

Now let $z = (\pi_3, N_3)$ then note the new logged likelihoods of $x+z$ and $x+y$ are the following

$$[\pi_1(1-q)N_1 + \pi_3(1-q)N_3]\ln(\frac{\sigma_g}{\sigma_b}) + [(1+\pi_1(q-1))N_1 + (1+\pi_3(q-1))N_3]\ln(\frac{1-\sigma_g}{1-\sigma_b})$$
$$\geq [\pi_2(1-q)N_2 + \pi_3(1-q)N_3]\ln(\frac{\sigma_g}{\sigma_b}) + [(1+\pi_2(q-1))N_2 + (1+\pi_3(q-1))N_3]\ln(\frac{1-\sigma_g}{1-\sigma_b}).$$

Note separability holds, and therefore, the updating rule is a monotonic function of the likelihood ratio.

Inertial Updating: Dominiak et al. (2023) follow a long line of literature which tries to model updating via a minimization problem involving the prior, likelihoods and posterior (Jaynes, 1957; Good et al., 1963; Williams, 1980; Shore and Johnson, 1980; Caticha and Giffin, 2006; Zhu et al., 2014). While the literature traditionally focused on Bayesian updating, Dominiak et al. (2023) contribute by showing it can be used to study non-Bayesian updating, and more importantly give it behavioral foundations. Their updating rule can be rewritten as

$$p(g|s) = \frac{g(p(g))f(\ell_g(s))}{g(p(g))f(\ell_g(s)) + g(1-p(g))f(\ell_b(s))}.$$

As f and g have flexible functional forms, the posterior need not be increasing in the likelihood ratio. The generality of this representation is due to the authors' commitment to a simple axiomatization, and the paper offers several special cases that satisfy monotonicity in likelihood ratio. In private conversation, the authors have shared they also strongly agree with monotonicity being a weak property.

B.2 Non-binary qualities with known accuracy is equivalent to binary quality with unknown accuracy

I show in this section that relaxing the assumption that the mental model has binary states, with known signal likelihoods, produces the same behavioral predictions as holding the binary state assumption and relaxing instead that signals likelihoods are known.

I say that $\succeq$ has a non-binary Bayesian expected utility representation with known accuracy if there is a set of qualities $q \in Q$, a positive and non-zero utility assigned to each quality $u(q)$, and for each quality a likelihood over signals of each type $\sigma_q = (\sigma_{q,1}, .., \sigma_{q,T})$, a prior $p(q)$ over qualities, such that

$$x \succeq y \text{ if and only if } \sum_Q u(q)p(q|x) \geq \sum_Q u(q)p(q|y).$$

This can be rewritten as follows

$$\sum_Q u(q)\frac{p(q)p(x|q)}{p(x)} \geq \sum_Q u_q \frac{p(q)p(y|q)}{p(y)},$$

then I write out the likelihoods,

$$\sum_Q u(q)p(q)\frac{\prod_{t\in T}\sigma_{q,t}^{x_t}}{p(x)} \geq \sum_Q u(q)p(q)\frac{\prod_{t\in T}\sigma_{q,t}^{y_t}}{p(y)},$$

and I expand the denominator,

$$\sum_Q u(q)p(q)\frac{\prod_{t\in T}\sigma_{q,t}^{x_t}}{\sum_Q p(q')\prod_{t\in T}\sigma_{q',t}^{x_t}} \geq \sum_Q u(q)p(q)\frac{\prod_{t\in T}\sigma_{q,t}^{y_t}}{\sum_Q p(q')\prod_{t\in T}\sigma_{q',t}^{y_t}}.$$

I show that if $\succeq$ has the above representation, then it also has a binary quality representation with accuracy uncertainty with Bayesian updating and expected utility.

Then a Bayesian expected utility maximizer with a set $q \in Q$ of potential accuracies, distribution p_g, p_b over accuracies given quality, and $p(g)$ priors behave as follows. Note for each accuracy q, I denote the vector by $\sigma_q = (\sigma_{q,1}, .., \sigma_{q,T})$. Note that $\frac{p(g)p_g(q)}{p(g)p_g(q)+p(b)p_b(q)} = p(g|q)$. First note now the DM chooses based on posterior therefore

$$x \succeq y \text{ if and only if } p(g|x) \geq p(g|y).$$

This can be written as follows

$$\sum_Q p(g,q|x) \geq \sum_Q p(g,q|y),$$

and then transformed by Bayesian updating

$$\sum_Q \frac{p(g,q)p(x|g,q)}{\sum_Q[p(g,q')+p(b,q')]p(x|q')} \geq \sum_Q \frac{p(g,q)p(y|g,q)}{\sum_Q[p(g,q')+p(b,q')]p(y|q')},$$

and writing out the likelihoods,

$$\sum_Q \frac{p_g(q)p(g)\prod_{t\in T}\sigma_{q,t}^{x_t}}{\sum_Q[p_g(q')p(g)+p_b(q')p(b)]\prod_{t\in T}\sigma_{q',t}^{x_t}} \geq \sum_Q \frac{p_g(q)p(g)\prod_{t\in T}\sigma_{q,t}^{y_t}}{\sum_Q[p_g(q')p(g)+p_b(q')p(b)]\prod_{t\in T}\sigma_{q',t}^{y_t}}.$$

To show the equivalence of the two representations, suffice to show that first for a given $u(q)$ and $p(q)$, we can find p_g, p_b and $p(g)$ such that $u(q)p(q) = p_g(q)p(g)$ and $p(q) = p(g)p_g(q) + p(b)p_b(q)$. And then show that for a given p_g, p_b and $p(g)$ we can find $u(q)$ and $p(q)$ where the equations hold.

Start with fixed $p(q)$ and $u(q)$, note that since the utility is linear, we can normalize $u(q)$ such that $\sum_Q u(q)p(q) \in (0,1)$ and all terms are positive, which implies all $u(q) \in (0,1)$. Set $p(g) = \sum_Q u(q)p(q) \in (0,1)$ and set $p_b(q)p(b) = p(q)[1-u(q)]$. Then $p_g(q) = \frac{u(q)p(q)}{\sum_Q u(q)p(q)}$ will ensure $u(q)p(q) = p_g(q)p(g)$ and sum up to 1 as desired so it is a probability distribution. Note $p(b) = 1 - p(g)$ then setting $p_b(g) = \frac{[1-u(q)]p(q)}{1-\sum_Q u(q)p(q)} \in (0,1)$ ensures the second equation holds while making sure it sums to one. The other direction is analogous.

C Online Appendix: Additional Tables

Table 14: Opting to Learn Given Sample Size Neglect

			% of Subjects Who Opt to Learn			
	% of Sample Size Neglect		Sample Size Neglect		No Sample Size Neglect	
Choice Task	Full	Subsample	Full	Subsample	Full	Subsample
0/4 vs 10	23%	42%	27%	26%	31%	27%
1/4 vs 10	39%	64%	22%	15%	28%	25%
2/4 vs 10	28%	34%	27%	18%	22%	18%
3/4 vs 10	58%	72%	19%	13%	28%	27%
4/4 vs 10	42%	71%	11%	10%	24%	14%
0/4 vs 25	39%	66%	21%	21%	37%	40%
1/4 vs 25	38%	54%	19%	13%	29%	30%
2/4 vs 25	38%	68%	20%	11%	33%	34%
3/4 vs 25	48%	69%	16%	10%	29%	28%
4/4 vs 25	42%	76%	12%	11%	27%	28%

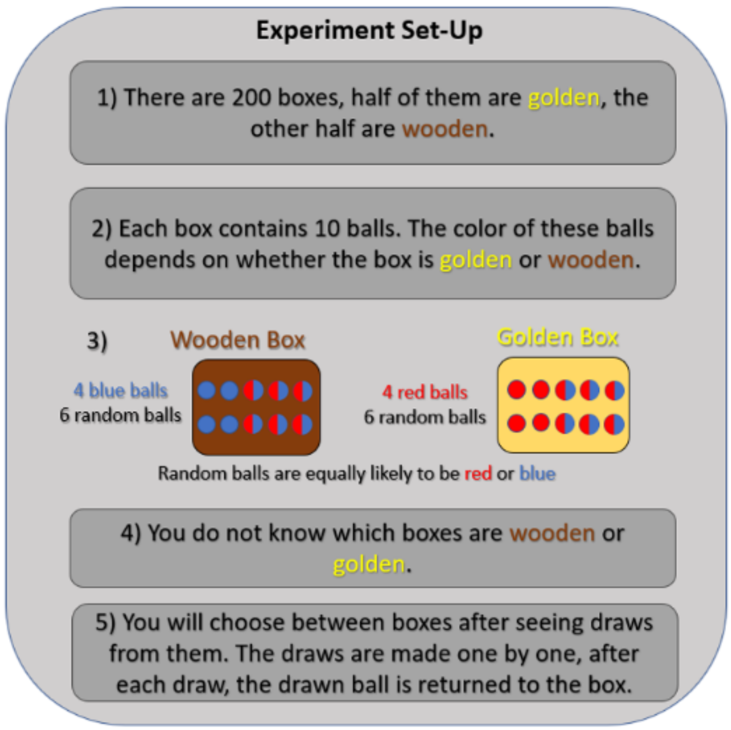
D Online Appendix: Experimental Instructions

D.1 Instructions

Experimental Instructions

In this experiment, you will be asked to choose between boxes. There are 200 boxes, 100 of them are golden, and another 100 are wooden. Each box also contains a number of balls of different colors, with the composition specified below.

Below is a graphical illustration of the experimental set-up:



Payment

In this experiment, you will be tasked to choose between boxes. One of the tasks has already been randomly selected by the computer for payment. If the box you chose for that task is a golden box, then you will earn a bonus payment of $10, if it is a wooden box, then you will not earn a bonus payment. Since you do not know which task was selected for payment, you should choose for each task as if it were the one chosen for payment. At the end of the experiment, the task chosen for payment will be revealed to you as well as the type of box you chose for that task.

Your payment is composed of two components:

1. You will be paid $4 for completing the experiment.
2. You will additionally earn $10 if your chosen box is golden in a randomly selected task.

The next button takes you to an example of a choice task and some comprehension questions before the experiment.

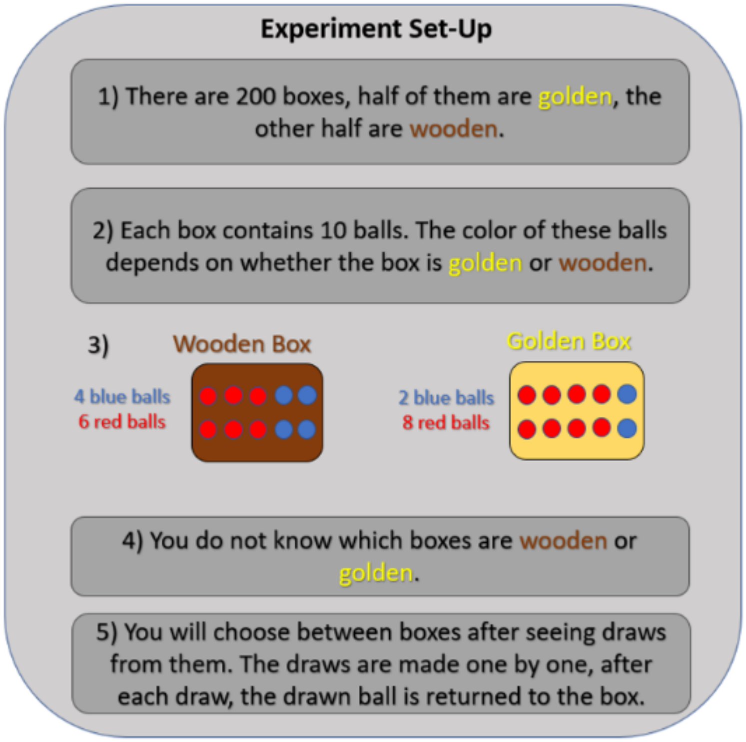
Next

Figure 10: Symmetric Treatment Instructions

Experimental Instructions

In this experiment, you will be asked to choose between boxes. There are 200 boxes, 100 of them are golden, and another 100 are wooden. Each box also contains a number of balls of different colors, with the composition specified below.

Below is a graphical illustration of the experimental set-up:



Payment

In this experiment, you will be tasked to choose between boxes. One of the tasks has already been randomly selected by the computer for payment. If the box you chose for that task is a golden box, then you will earn a bonus payment of $5, if it is a wooden box, then you will not earn a bonus payment. Since you do not know which task was selected for payment, you should choose for each task as if it were the one chosen for payment. At the end of the experiment, the task chosen for payment will be revealed to you as well as the type of box you chose for that task.

Your payment is composed of two components:

1. You will be paid $2.5 for completing the experiment.
2. You will additionally earn $5 if your chosen box is golden in a randomly selected task.

The next button takes you to an example of a choice task and some comprehension questions before the experiment.

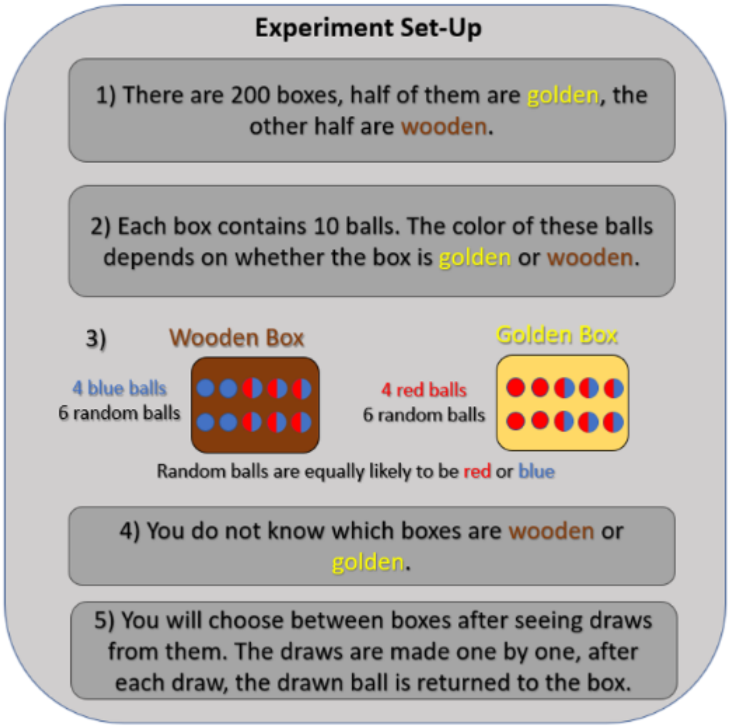
Next

Figure 11: Asymmetric Treatment Instructions

Experimental Instructions

In this experiment, you will be asked to choose between boxes. There are 200 boxes, 100 of them are golden, and another 100 are wooden. Each box also contains a number of balls of different colors, with the composition specified below.

Below is a graphical illustration of the experimental set-up:



Payment

In this experiment, you will be tasked to choose between boxes. One of the tasks has already been randomly selected by the computer for payment. If the box you chose for that task is a golden box, then you will earn a bonus payment of $5, if it is a wooden box, then you will not earn a bonus payment. Since you do not know which task was selected for payment, you should choose for each task as if it were the one chosen for payment. At the end of the experiment, the task chosen for payment will be revealed to you as well as the type of box you chose for that task.

Your payment is composed of two components:

1. You will be paid $2.5 for completing the experiment.
2. You will additionally earn $5 if your chosen box is golden in a randomly selected task.

The next button takes you to an example of a choice task and some comprehension questions before the experiment.

Next

Figure 12: Correlated Treatment Instructions

Introduction

Hover to see the experimental set-up.

Throughout the experiment, you will have to choose between boxes. In this example, you are presented with two boxes. Each of the two boxes were randomly selected from the pool of 200 boxes. If the box you choose is golden and this choice is randomly selected by the computer, then you will receive an additional $5. Here is how boxes A and B differ:

1. The computer drew 5 balls from Box A one at a time, returning each ball to the box after it was drawn. Of the 5 balls drawn, 3 were red (and the other 2 were blue).
2. The computer will draw 10 balls from Box B and your choice can be based on the outcome of the draws.

Note that the more red balls drawn from a box, the more likely it is golden rather than wooden. So if you pick Box B when 6 red balls out 10 balls are drawn from it, you should still pick it for 7, 8, 9 and 10 red balls out of 10. The interface below is implemented to automatically make these selections for you.

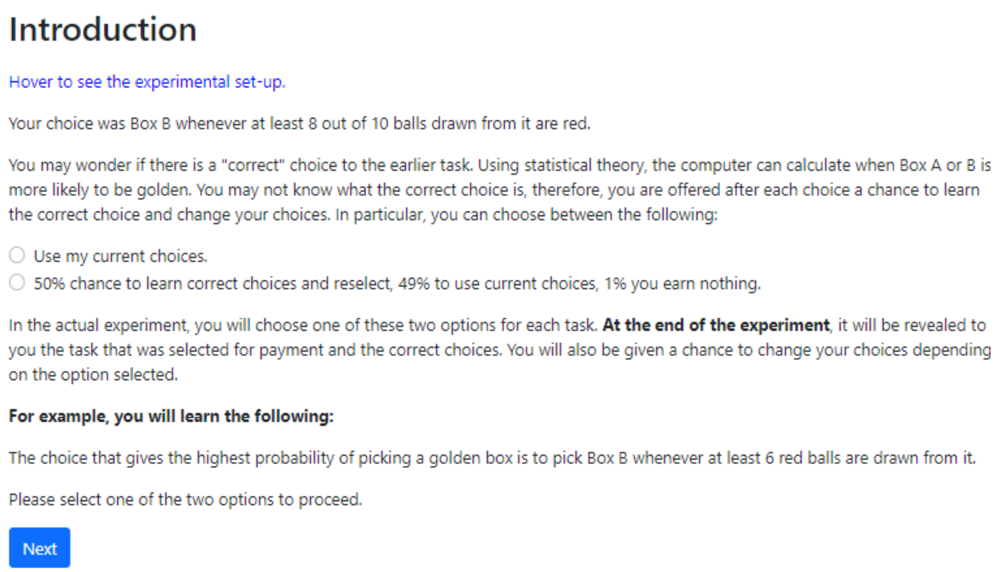
The interface will automatically fill out your choices in the following way. If you click on the right column for a row, say you choose Box B when 4 red balls out of 10 are drawn from it , then for all the rows above this row the computer selects Box A and for all the rows below it selects Box B. Similarly, if you choose Box A when Box B draws 7 red balls out of 10 and click on the left column for that row, then for all the rows above, the computer selects Box A and for all the rows below it selects Box B. You can also refresh the page to start over anytime.

Choose Box A		Choose Box B
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **0 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **1 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **2 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **3 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **4 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **5 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **6 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **7 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **8 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **9 out of 10** balls drawn are red.
Box A - **3 out of 5** balls drawn were red.	○ ○	Box B - if **10 out of 10** balls drawn are red.

In order to proceed, please refresh the page and make the hypothetical selections (underlined) of the previous paragraph in the stated order to see for yourself how the interface works. The button will be enabled after you perform the selections of the above paragraph.

Next

Figure 13: Payment and Choice Example



Introduction

Hover to see the experimental set-up.

Your choice was Box B whenever at least 8 out of 10 balls drawn from it are red.

You may wonder if there is a "correct" choice to the earlier task. Using statistical theory, the computer can calculate when Box A or B is more likely to be golden. You may not know what the correct choice is, therefore, you are offered after each choice a chance to learn the correct choice and change your choices. In particular, you can choose between the following:

- Use my current choices.
- 50% chance to learn correct choices and reselect, 49% to use current choices, 1% you earn nothing.

In the actual experiment, you will choose one of these two options for each task. **At the end of the experiment**, it will be revealed to you the task that was selected for payment and the correct choices. You will also be given a chance to change your choices depending on the option selected.

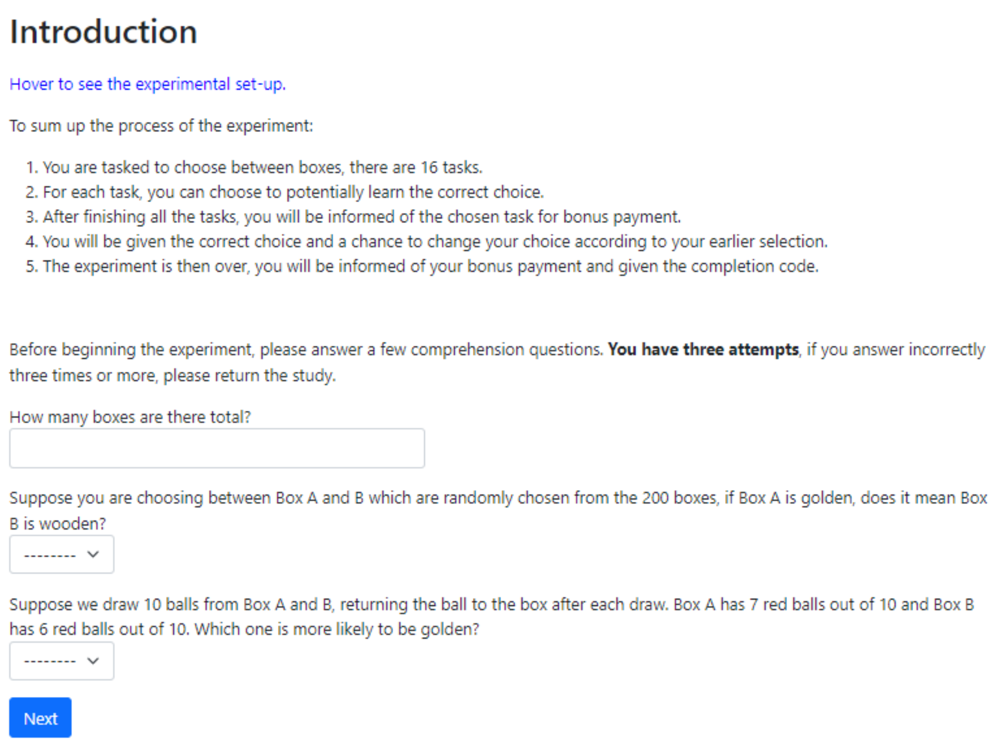
For example, you will learn the following:

The choice that gives the highest probability of picking a golden box is to pick Box B whenever at least 6 red balls are drawn from it.

Please select one of the two options to proceed.

Next

Figure 14: Confidence Elicitation Instructions



Introduction

Hover to see the experimental set-up.

To sum up the process of the experiment:

1. You are tasked to choose between boxes, there are 16 tasks.
2. For each task, you can choose to potentially learn the correct choice.
3. After finishing all the tasks, you will be informed of the chosen task for bonus payment.
4. You will be given the correct choice and a chance to change your choice according to your earlier selection.
5. The experiment is then over, you will be informed of your bonus payment and given the completion code.

Before beginning the experiment, please answer a few comprehension questions. **You have three attempts**, if you answer incorrectly three times or more, please return the study.

How many boxes are there total?

Suppose you are choosing between Box A and B which are randomly chosen from the 200 boxes, if Box A is golden, does it mean Box B is wooden?

Suppose we draw 10 balls from Box A and B, returning the ball to the box after each draw. Box A has 7 red balls out of 10 and Box B has 6 red balls out of 10. Which one is more likely to be golden?

Next

Figure 15: Comprehension Check

D.2 Choice Examples

Choice

Hover to see the experimental set-up.

You are offered a choice between two boxes, A and B. Each of these boxes were randomly picked from the 200 boxes. You will be paid $5 if the box you pick is golden and the computer has selected this task for payment. Box A and B were randomly picked the following way:

1. Box A was picked as follows:
   1. The computer drew 4 balls from each of the 200 boxes. They were drawn one at a time, returning each ball to the box after it was drawn.
   2. 52 out 200 boxes had 2 out 4 red balls drawn from them (2 other balls were blue).
   3. Box A was randomly selected from these 52 boxes.

2. Box B was picked as follows:
   1. The computer randomly picked Box B from the 200 boxes.
   2. The computer will draw 10 balls from Box B, one at at time and returning the drawn ball to the box.
   3. You can make your choice based on the number of red balls that are drawn from B.

Choose Box A	Choose Box B
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **0 out of 10 (0%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **1 out of 10 (10%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **2 out of 10 (20%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **3 out of 10 (30%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **4 out of 10 (40%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **5 out of 10 (50%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **6 out of 10 (60%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **7 out of 10 (70%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **8 out of 10 (80%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **9 out of 10 (90%)** balls drawn are red.
Box A - **2 out of 4 (50%)** balls drawn were red.	Box B - if **10 out of 10 (100%)** balls drawn are red.

You may wonder if there is a "correct" choice to this task. Using statistical theory, the computer can calculate when Box A or B is more likely to be golden. You may not know what the correct choice is, therefore, you are offered after each choice a chance to learn the correct choice and change your choices. In particular, you can choose between the following:

Choose an option:

Next

Figure 16: Comparative 4 vs 10 Balls Example

Choice

Hover to see the experimental set-up.

You are offered a choice between two boxes, A and B. Each of these boxes were randomly picked from the 200 boxes. You will be paid $5 if the box you pick is golden and the computer has selected this task for payment. Box A and B were randomly picked the following way:

1. Box A was picked as follows:
   1. The computer drew 4 balls from each of the 200 boxes. They were drawn one at a time, returning each ball to the box after it was drawn.
   2. 25 out 200 boxes had 0 out 4 red balls drawn from them (4 other balls were blue).
   3. Box A was randomly selected from these 25 boxes.

2. Box B was picked as follows:
   1. The computer randomly picked Box B from the 200 boxes.
   2. The computer will draw 25 balls from Box B, one at at time and returning the drawn ball to the box.
   3. You can make your choice based on the number of red balls that are drawn from B.

Choose Box A		Choose Box B
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **0 out of 25 (0%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **1 out of 25 (4%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **2 out of 25 (8%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **3 out of 25 (12%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **4 out of 25 (16%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **5 out of 25 (20%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **6 out of 25 (24%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **7 out of 25 (28%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **8 out of 25 (32%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **9 out of 25 (36%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **10 out of 25 (40%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **11 out of 25 (44%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **12 out of 25 (48%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **13 out of 25 (52%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **14 out of 25 (56%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **15 out of 25 (60%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **16 out of 25 (64%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **17 out of 25 (68%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **18 out of 25 (72%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **19 out of 25 (76%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **20 out of 25 (80%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **21 out of 25 (84%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **22 out of 25 (88%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **23 out of 25 (92%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **24 out of 25 (96%)** balls drawn are red.
Box A - **0 out of 4 (0%)** balls drawn were red.	○ ○	Box B - if **25 out of 25 (100%)** balls drawn are red.

You may wonder if there is a "correct" choice to this task. Using statistical theory, the computer can calculate when Box A or B is more likely to be golden. You may not know what the correct choice is, therefore, you are offered after each choice a chance to learn the correct choice and change your choices. In particular, you can choose between the following:

Choose an option:

Next

Figure 17: Comparative 4 vs 25 Balls Example

Choice

Hover to see the experimental set-up.

You are offered a choice between two boxes, A and B. You will be paid $5 if the box you pick is golden and the computer has selected this task for payment. Box A and B are different in the following way:

1. Box A has a 25% chance of being golden.

2. Box B was randomly selected from the 200 boxes.
    1. The computer will draw 4 balls from Box B, one at at time and returning the drawn ball to the box.
    2. You can make your choice based on the number of red balls that are drawn from B.

Choose Box A		Choose Box B
Box A - **25%** chance of being golden.	○ ○	Box B - if **0 out of 4 (0%)** balls drawn are red.
Box A - **25%** chance of being golden.	○ ○	Box B - if **1 out of 4 (25%)** balls drawn are red.
Box A - **25%** chance of being golden.	○ ○	Box B - if **2 out of 4 (50%)** balls drawn are red.
Box A - **25%** chance of being golden.	○ ○	Box B - if **3 out of 4 (75%)** balls drawn are red.
Box A - **25%** chance of being golden.	○ ○	Box B - if **4 out of 4 (100%)** balls drawn are red.

You may wonder if there is a "correct" choice to this task. Using statistical theory, the computer can calculate when Box A or B is more likely to be golden. You may not know what the correct choice is, therefore, you are offered after each choice a chance to learn the correct choice and change your choices. In particular, you can choose between the following:

Choose an option:

Next

Figure 18: Belief Elicitation 4 Balls Example

D.3 Pre-Payment and Payment

Thought Process

Before proceeding to showing you the task the computer selected for payment (and the correct choices), we'd like to know your thought process behind your choices. In particular, it would help us if you answered the following questions.

In the experiment, you were offered a chance to learn the correct choice. Sometimes you may have chosen to not use it. Which of the following best describes why you chose not to use it?

○ I did not use it because I knew the correct choice.
○ I did not use because of the 1% chance of not earning a bonus payment.
○ I always used it.

Overall, how sure are you that you made the right choices?

○ I know my choices were close to correct (for most tasks).
○ I do not know if my choices were close or not to correct (for most tasks).

In the experiment, how would you describe your choice when choosing between two boxes that both had balls drawn from them?

○ I relied mostly exclusively on the percentage of red balls drawn.
○ I used the percentage and weighted it by the total number of balls drawn.
○ I tried to estimate how likely each box was golden independently.
○ I used something else entirely.

Next

Figure 19: Unincentivized Confidence Elicitation

PrePayment

Hover to see the experimental set-up.

The task that was selected for payment was the following. We repeat the instructions below:

You are offered a choice between two boxes, A and B. Each of these boxes are randomly picked from the 200 boxes. You will be paid $10 if the box you chose is golden and if this choice is selected for payment. Box A and B are different in the following way:

1. We drew 4 balls from Box A, **2 out of 4 (50%)** are red.
2. We will draw 25 balls from Box B and your choice can be based on the outcome of the draws.

In this task, you chose box B if it drew at least **7 out of 25** red balls. And when given the chance to learn the optimal choice, you chose to have a 50% chance to learn the correct choices, 49% to use current choices, 1% you earn nothing. **You got the option to learn the correct choice and can reselect.**

Please enter the smallest number of red balls Box B needs to draw for you to choose it:

0

Below are the optimal choices for the different outcomes of Box A.

The draw from Box A	When to choose Box B
0 out 4 are red balls (the rest are blue).	If it draws at least 11 out of 25 red balls.
1 out 4 are red balls (the rest are blue).	If it draws at least 12 out of 25 red balls.
2 out 4 are red balls (the rest are blue).	If it draws at least 13 out of 25 red balls.
3 out 4 are red balls (the rest are blue).	If it draws at least 14 out of 25 red balls.
4 out 4 are red balls (the rest are blue).	If it draws at least 15 out of 25 red balls.

Next

Figure 20: Learning from the Confidence Elicitation Mechanism Example

Payment

You are done! Below we summarize your earnings:

In this task, you chose box B if it drew at least **7 out of 25** red balls. In the end, Box B drew **24 out of 25** red balls, which means you chose **Box B**. The box was **wooden** , you will not earn a bonus payment.

Additionally, the completion code for prolific is CKZ0D6QO.

If you have any questions, please reach us via Prolific.

Figure 21: Final Payment Example