The Structure of Sequential Updating

Kim Sarnoff *

February 24, 2026

updated often – please click here for latest version

Abstract Many real-world inference problems unfold over time: employers learn about ability across tasks, consumers evaluate products through repeated use, and policymakers revise beliefs as new data arrive. Yet despite its ubiquity, research on dynamic updating has largely focused on a single implication of Bayesian reasoning: order independence. This paper experimentally tests a broader set of restrictions implied by Bayes' rule, emphasizing both order independence and the previously unexamined property of *prior sufficiency*: the principle that the most recent posterior should serve as a sufficient statistic for past information. In a multi-period updating experiment with a rich set of parameters, participants repeatedly revise beliefs after receiving signals of varying strength and structure. Three main results emerge. First, only roughly a third display order dependence, overreacting to conflicting signals. Second, violations of prior sufficiency are widespread: beliefs formed sequentially tend to grow more extreme, and models assuming prior sufficiency, such as Grether (1980), fit poorly beyond the first update. Finally, the data indicate that participants process signals in aggregate, explaining prior sufficiency violations.

Thank you to Nava Ashraf, Alexia Delfino, Alex Imas, Alessandro Lizzeri, Irene Solmone, Charlie Sprenger, Sevgi Yuksel, and participants in Princeton Microeconomic Theory Lunch for helpful comments, and an extra special thank you to Ilyana Kuziemko and Leeat Yariv.

*Contact: ksarnoff@princeton.edu

1 Introduction

Many real-world inference problems unfold gradually, as information accumulates over time. Investors update beliefs about asset quality with each earnings report, consumers learn about products through repeated use, teachers infer students' ability from successive evaluation, and policymakers revise their views of the economy as new statistics arrive. Bayesian updating provides the normative benchmark for how beliefs should adjust in light of new evidence and underpins much of decision theory under uncertainty.

A vast experimental literature documents systematic deviations from Bayesian updating, typically using one-shot settings where individuals receive a prior, observe a single signal, and then report a posterior belief. The smaller literature on dynamic updating has focused primarily on how posterior beliefs depend on the order in which signals arrive. This paper experimentally studies a broader set of properties of sequential updating, across varying sets of parameters. I consider not only the traditional focus of the literature — whether updating is *order independent* — but also whether updating adheres to *prior sufficiency* — the principle that, when information arrives sequentially, the previous posterior serves as a sufficient statistic for updating from the next signal.

Overall, participants in my experiment broadly conform to the core features of Bayesian updating. They adjust their beliefs in the correct direction when new information arrives, and their posteriors are, on average, close to the statistical benchmark. However, I identify three key findings regarding deviations from Bayesian reasoning. First, although order dependence is substantial in magnitude, it appears for only one-quarter to one-third of participants and follows a distinct pattern — an overreaction to the most recent signal when it conflicts with the prior. Second, unlike order independence, prior sufficiency fails across the population: posteriors are systematically and substantially more extreme when formed through sequential updating than when based on a sufficient statistic summarizing the same information. Consistent with this empirical finding, variants of the canonical Grether (1980) model — which assumes prior sufficiency — fail to capture behavior. Third, I find evidence that subjects consider sequences in the aggregate, explaining failures of prior sufficiency.

In the experiment, participants are tasked with guessing an underlying state. They receive four conditionally independent and identically distributed (i.i.d.) signals sequentially and report their posterior belief after each one. The design mirrors real-world settings in which information arrives incrementally. Across treatments, I vary the precision of the signals, and the prior distribution over states.

The inference tasks that a subject faces are designed to create within person tests of

two key properties implied by Bayesian updating in dynamic environments. The first is order independence: when signals are conditionally i.i.d., their order of arrival should not affect posterior beliefs. The second is prior sufficiency: at any point in time, the posterior from a subset of signals fully summarizes the information relevant for future updating, so the specific composition or order of those signals should no longer matter.

Violations of these properties have far-reaching implications. When order independence fails, the sequencing of information itself can shape beliefs — lawyers may sway jurors through the order in which evidence is presented, and news outlets may influence readers simply by rearranging stories. A failure of prior sufficiency, in contrast, means that identical information can be interpreted differently depending on prior exposure — an incumbent and a challenger making the same statement may be judged unequally, or an employer may evaluate a long-time employee and a new hire with identical credentials differently. To my knowledge, there are no papers with similar within person variation in sequence order for more than two signals, and only two papers that consider prior sufficiency in any way (Möbius et al., 2022; Raymond & Wittrock, 2024).

To contrast the particular implications of each violation more sharply, consider a manager who must learn about a worker's ability. Anticipating our experimental setup, there are two ability types: $H(igh)$ and $L(ow)$. The manager has an initial prior about the probability the worker is H and updates this belief by observing the worker's performance on tasks. Each task results in a binary signal, h or l, where an H worker is more likely to send h and an L worker is more likely to send l.

Suppose the manager has two new hires today, called Worker $N(ew)1$ and Worker $N(ew)2$. The manager's prior about each worker comes from a report from his respective previous employer. The reports are identical, so that the manager has the same prior for both. In their first two tasks, both $N1$ and $N2$ send one h signal and one l signal, but in opposite orders — $N1$ sends an h first and $N2$ sends an l first. Order independence requires that the manager hold the same posterior about both workers at the end of the second task, since the initial prior and aggregate information received are the same.

Now, let us compare Worker $N1$ to a worker who has been employed for a year, called Worker $O(ld)$. The manager's prior about Worker O is formed endogenously through observing O's performance, where the prior for O is the posterior after the most recent task. So, $N1$ and O differ from each other in terms of the process by which the prior was generated — endogenously for Worker O and exogenously for Worker $N1$. Suppose that on $N1$'s first day, the manager's belief about Worker O is numerically identical to the prior about $N1$. Additionally, Worker O's subsequent signal is an h, matching $N1$'s first signal. Prior sufficiency requires the manager to update identically for both workers, since the

posterior should depend only on the prior value and the signal, and not on the way the prior was formed. This example clarifies how order independence and prior sufficiency are not coincident properties. The manager could update the same way about Worker $N1$ and $N2$, but not about Worker $N1$ and O, or vice versa.

The experimental design constructs counterfactuals to test both of these properties within subject. The basis of the design is a standard ball-and-urn setup with two states and binary symmetric signals. Subjects complete two different inference tasks: a sequential task and a one-shot task. In the sequential task, subjects face a dynamic problem where they observe four signals incrementally, reporting a posterior after each. In the one-shot task, subjects face a static problem, where they receive an exogenous prior, and report a posterior for both possible signal realizations via strategy method. I collect one-shot updates for priors at five percentage point intervals and for both signals. Treatments vary between subjects the precision q of the signal, and the initial prior in the sequential task. The two baseline treatments use a precision of 0.6 or 0.8, with a uniform initial prior.

The sequential task provides the variation needed to test order independence. This test is fairly straightforward to construct. I gave each subject multiple sequences where three signals are of one type and the fourth signal conflicts. Across sequences, I vary within subject when the conflicting information appears: at time 1, 2, and 3 for subsequences of length 3, and at time 2, 3, and 4 for full sequences. This design corresponds to the comparison of Workers $N1$ and $N2$ from our example, who have the same prior and same aggregate information but differently ordered signals.

Testing prior sufficiency is more challenging. After observing the first signal, each subject solves an endogenously determined problem — their posterior becomes their prior for the next update, and this prior reflects their subjective processing of information. The experimenter controls the objective information subjects receive, but only indirectly controls the problem the subject solves. Contrast this with the test of order independence, where the experimenter always retains full control over the necessary variation (namely, signal order).

The experimental design overcomes this challenge by using the sequential and one-shot tasks in tandem. For each update in a sequence, I match to the one-shot task where the same subject updates from the same prior value with the same signal. The only difference is that in the sequential task, the prior is formed endogenously through the subject's own updating, while in the one-shot task, the prior is given exogenously. To analogize to our manager example, the update about Worker O corresponds to an update in the sequential task. The update about Worker $N1$ corresponds to the one-shot counterfac-

tual for this sequential update. The strength of this test is that it does not hinge on any assumption about the functional form of the updating rule. If updating adheres to prior sufficiency, then we will observe no difference between the two tasks, regardless of any other detail about the procedure being used. Because I collect updates for a grid of priors, I obtain a counterfactual mapping from prior, signal pairs to posteriors for almost all sequential reports.

I begin by documenting two patterns about the prevalence of order dependence and prior sufficiency. Order dependence occurs in a minority of the population, while prior sufficiency is violated uniformly. For order dependent subjects, reports diverge specifically between sequences where the conflicting signal is most recent and sequences where the conflicting information has already occurred. This behavior is correlated within person across sequences.

Using a mixture model to divide the population, I find it splits precisely on overreaction to conflicting information. Analyzing order effects for these two groups separately, we confirm that the majority exhibit no order effect in almost all cases. For the minority, beliefs are between 25.3 and 56.9 percentage points lower when the sequence ends in a conflicting signal. This is driven by the qualitative mistake of reporting a posterior in favor of one state when the overall evidence favors the other. In other words, subjects respond to the conflicting signal in the right direction, but the magnitude is too large. There is essentially no gap between the sequences where the conflicting signal has already occurred.

I then examine prior sufficiency separately by type. There are two reasons for this. The first is that the majority grasp at least one feature of Bayesian reasoning (order independence), stacking the cards against us in terms of identifying additional violations. The second is that order dependence might in fact follow from prior sufficiency, an explanation that the literature has not evaluated. To see how, let us return to the comparison of the two new workers, $N1$ and $N2$, from our manager example. After task 1, the belief about $N1$ and $N2$ should be different: different information has been received. If the manager adheres to prior sufficiency but updates in a way that depends on the prior value, then the posteriors may remain different after task 2, even though the aggregate information is now the same.

I find that both types fail the test for prior sufficiency. Order independent subjects react more to a non-conflicting signal in the sequence and less to a conflicting signal in the sequence, relative to the one-shot counterfactual. The result is that reports are more extreme in the sequential task for all sequences. Depending on the treatment and time period, the gap in reports is as large as 15 percentage points, which is on the order of

additional signal.

The minority group's behavior cannot be explained by prior sufficiency. Estimates are noisier, given the sample is smaller, but we observe two key behaviors after the 4th signal. For the sequence with conflicting information at time 4, reports are 12 percentage points lower in the sequential task than in the one-shot task. This establishes that overreaction to a conflicting signal is specific to the sequence. Concomitantly, for the sequence with conflicting information at time 3, reports in the sequential task are between 5 and 13 percentage points higher. This shows that the pattern of correction immediately after a conflicting signal is also sequence specific.

A natural question is whether prior sufficiency fails because of the presentation of the problem — namely, that individuals cannot recognize that a prior technically represents the same information regardless of its source — or because of variation in the process by which beliefs are generated — an individual may treat an endogenous belief that they have formed differently from an exogenous one provided to them.

I address this with additional treatments that test prior sufficiency with respect to a *sequence*, rather than a single signal. This test compares the uniform treatment with precision q to a paired treatment with the same precision and a *non-uniform prior* equal to q. The core idea is that the non-uniform treatment presents sequential problems, rather than one-shot problems, that are informationally equivalent to the uniform sequences.

In the uniform treatment, the Bayesian posterior after observing the first signal equals q — therefore, the non-uniform prior is informationally equivalent to one signal. This property allows me to take any uniform sequence and construct an equivalent non-uniform version by removing the first signal and shifting the signals for times 2 through 4 up to times 1 through 3. The only objective difference between the pair of sequences is whether the first piece of information enters through the prior or through an observed signal.

This design generates two different tests of prior sufficiency. The first is an 'exact' test. Subjects in the uniform treatment who report q after the first signal have an endogenous prior at time 2 that is exactly equal to the exogenously set prior for a non-uniform subject. Under prior sufficiency, we should observe the same distribution of reports between these uniform subjects and the non-uniform subjects at each subsequent period. For the weaker signal precision of 0.6, 60% of reports in the uniform treatment after the 1st signal equal the Bayesian posterior, making this exact test viable.

For the stronger precision of 0.8, only 30% of initial reports equal 0.8, precluding the exact test on any reasonable sample size. However, we can conduct a second distribution-based test that exploits a result from Chan (2025): under a broad class of Grether (1980) type updating rules where the prior is power distorted and the signal distorted according

to any function, the posterior distribution for the non-uniform treatment should remain above the distribution for the uniform treatment.

On both tests, prior sufficiency continues to fail. For $q = 0.6$, when we restrict to reports from sequences where the first posterior is Bayesian, beliefs are more extreme than in the corresponding non-uniform sequence. This is the same qualitative pattern as in the one-shot comparison. For $q = 0.8$, we observe a reversal in the ordering of distributions: reports in the non-uniform treatment initially exceed reports in the uniform treatment for the equivalent sequence, but by the fourth signal, reports in the uniform exceed reports in the non-uniform. We observe this reversal for the weaker signal as well, where 80% of time 1 reports are less than or equal to 0.6.

With these non-parametric results in hand, we turn to the matter of estimating updating rules parametrically. I consider the benchmark econometric model of Grether (1980), which implies that sequential updating follows an AR1 process. This AR1 structure is precisely the assumption of prior sufficiency — that beliefs depend only on the prior (the lagged posterior) and current signal. The results suggest that such a model will do a poor job of organizing the data. I estimate a variety of Grether parameterizations, using a simulated maximum likelihood approach to handle the endogeneity of the lagged posterior and to recover a distribution of parameters for assessing model fit. Consistent with the empirical findings, the structural estimation fails to closely match behavior in the sequence. In particular, the Grether model underestimates the frequency of the modal report, and mismatches the distribution, typically by overestimating the variance of reports.

In the final section, I exploit additional design features to provide suggestive evidence of what drives failures of prior sufficiency. First, I make use of variation in the timing of equivalent sets of signals. Due to the structure of the problem, the Bayesian posterior depends only on the initial prior and the net count of signals. For example, a sequence with two 1 signals and one 0 signal is informationally equivalent to a single 1 signal. The sequences are designed so subjects encounter the same reduced set of information within the first two periods and within the second two periods. What changes across these two cases is that the prior may be different (endogenously) and more signals have accumulated without changing the objective posterior. We can control for the former through the one-shot reports. Any residual indicates that signal accumulation matters independently. Even when accounting for one-shot reports, beliefs are 4 to 10 percentage points higher in the second half of the sequence, when more signals have accumulated. That reports increase with signal accumulation suggests people may use a frequency-based procedure, forming different predictions as the set of signals changes.

Understanding prior sufficiency is important because it has distinct theoretical, econometric, and practical implications relative to order independence. From a theoretical perspective, consider someone who violates prior sufficiency but is order independent. Then, within any context, they follow a consistent rule with a known structure: a mapping from a prior belief and signal to a posterior. In this case, our goal might be to develop models that explain why a given prior, signal pair produces a particular posterior value. If individuals fail prior sufficiency, we require alternative theories about which variables enter this mapping altogether.

For empirical research, the coincidence of these biases impacts how we should generalize findings across different experimental settings. The majority of belief updating experiments use one-shot problems. If updating is not prior sufficient, one would have to determine whether any static experiment could mimic the additional state variables at play in sequential updating. If not, then findings from one-shot problems may be inapplicable for guiding work on sequential contexts.

As discussed earlier, there are also econometric implications. If individuals update in a prior sufficient way, beliefs follow an AR(1) process. The benchmark empirical approach in the updating literature, Grether (1980), implicitly assumes this structure when extended beyond one period. These models can accommodate various forms of miscalibration that generate order or time dependence, but they are misspecified if prior sufficiency fails. In such cases, we would need to include a different state variable that encodes features of the history beyond the current belief.

Finally, there are practical implications for policy design. Consider a policymaker designing an information intervention, such as a public health campaign about vaccine efficacy. The optimal way to deliver information depends on which properties the population satisfies. If the population exhibits order dependence, changing the introduction of information will matter for final beliefs, potentially favoring interventions that deliver all information simultaneously. If the population is order independent but not prior sufficient, sequential delivery might be preferable — beliefs formed endogenously may be better calibrated or more robust than those formed from equivalent summary statistics. These biases will also affect how well a policymaker can extrapolate from past experience. For example, suppose the government is deciding whether a successful flu vaccination campaign that delivered information over months would be equally effective for COVID vaccination, where information has to be delivered immediately. If the population is not prior sufficient, it may be unclear which aspects of the flu strategy to apply.

Related literature This paper contributes to several strands of research on deviations from Bayesian updating. First, it contributes to the literature on biases specific to sequential updating. This body of work has predominantly focused on characterizing order dependence, with mixed evidence on the direction in which sequencing matters — some studies find recency bias and others find confirmation bias (see Benjamin (2019) and references therein). More recent papers consider the role of format in sequential updating, e.g. how people respond to retractions versus new information (Gonçalves et al., 2025), or how non-linearities affect updates (Agranov & Reshidi, 2024).

I will comment on three papers that are closest to this one. Kieren et al. (2025) study reaction to disconfirming information in sequential problems. Consistent with my results, they show that people overreact when a signal contradicts a prior streak, and then "cancel" with the subsequent signal, returning roughly to their initial belief. The within person design of this paper helps to identify several additional phenomena with respect to order effects: first, that they are heterogenous, and second that they are not driven by adherence to prior sufficiency.

Agranov and Reshidi (2024) focus on a different question from this paper: how people integrate two signals of different precisions. As in my treatments, they compare updating from a uniform prior to updating from an equivalent non-uniform prior with one additional signal. By extending the sequence from two signals to create more than one endogenous update, I obtain a different test from theirs: namely, whether people exhibit prior sufficient behavior at some point in a sequence.

Chan (2025) axiomatically characterizes the Grether (1980) updating rule and experimentally tests some of its implications, including order independence. They do not consider prior sufficiency. This paper has a different focus, both in terms of the properties studied — I consider prior sufficiency, while they do not — and variation in environments.

As far as I am aware, there are only two papers to study prior sufficiency empirically in any way. The first is Möbius et al. (2022). They test a variety of propositions on a panel of belief updates that they collect, one of which is prior sufficiency — their test is completely econometric, and checks whether higher order lags of beliefs are predictive of the current belief. In this paper, I not only have designed variation to identify prior sufficiency, but I address weaknesses of econometric approaches that use the lag directly. The second is Raymond and Wittrock (2024), who test whether individuals remember signals or beliefs when updating sequentially. This experiment shuts down memory by design.

Second, this paper speaks to biases predominantly studied in the context of one-shot

updating tasks. The results suggest that base rate neglect, one of the most commonly studied biases related to the prior (Kahneman and Tversky (1973); Esponda et al. (2024)), may not be as relevant in sequential problems. Namely, the gap between the uniform and non-uniform treatments persists over time, suggesting that subjects do not ignore the prior even as signals accumulate.

Finally, this paper contributes to the literature on description versus experience in decision-making. One interpretation of the one-shot versus sequential comparison is that one-shot problems describe the inference task while sequential problems require learning through experience. Classic work in this area (e.g., Hertwig et al. (2004)) has documented systematic differences between decisions from description versus experience in risky choice settings. In contrast to work in risky choice, this experiment has no feedback, so a wedge between treatments cannot arise from learning the correct decision through experience.

The structure of the paper is as follows. Section 2 discusses the experimental design, Sections 3 and 4 present aggregate results on order independence and prior sufficiency with respect to a single signal, Section 5 examines heterogeneity in these two properties, Section 6 performs an alternative test of prior sufficiency with respect to a sequence, Section 7 evaluates the performance of the literature's standard empirical model, and Section 8 considers mechanisms behind observed failures of prior sufficiency. Section 9 concludes.

2 Design

Participants complete two inference tasks: a sequential inference task and a one-shot inference task. The basic inference problem is the same in both tasks, and is described to participants with a ball and urn setup.

The environment is as follows. There are two decks of cards, labeled Deck A and Deck B. Each deck contains 10 cards, where a card is one of two types — blue or orange. Deck A has majority blue cards and Deck B has majority orange cards.

A deck is drawn according to prior distribution $p(\text{Deck } A)$, denoted p. This is represented as a wheel, divided between the decks in proportion to their probabilities. The participant spins the wheel, and the deck where it stops is selected.

The objective is to predict the probability that each deck was drawn, given t signals. The participant receives signals by drawing cards at random with replacement from the selected deck. Signals are conditionally i.i.d. with precision q. That is, $p(\text{draw } blue \text{ card}|\text{Deck}A) = p(\text{draw } orange \text{ card}\,|\,\text{Deck } B) = q$. Let n_0 denote the number of orange cards observed, and

n_1 denote the number of blue cards observed. The Bayesian posterior probability of Deck A, given n_0 and n_1 is:

$$p(Deck\,A \mid n_0,\, n_1) = \frac{q^{n_1}(1-q)^{n_0} \cdot p(Deck\,A)}{q^{n_1}(1-q)^{n_0} \cdot p(Deck\,A) + q^{n_0}(1-q)^{n_1} \cdot p(Deck\,B)}$$

Treatments vary the signal precision, which is common to both tasks, and the prior in the sequential task. The priors in the one-shot task are the same for all treatments. There are 4 parameterizations, randomized between subject:

Treatment	**Prior** p	**Precision** q
1	0.5	0.6
2	0.5	0.8
3	0.6	0.6
4	0.8	0.8

Treatments 1 and 2 have a uniform prior with $q \in \{.6,.8\}$. Treatments 3 and 4 have a non-uniform prior equal to the signal precision, where $q = p \in \{.6,.8\}$.

We explain the sequential inference task and then the modifications made for the one-shot inference task.

2.1 Sequential inference tasks

The participant completes 6 sequential inference tasks, where the first task is unincentivized practice and the subsequent 5 tasks are incentivized.[1] A task consists of four identical rounds, with the following steps:

1. Information: The participant draws one card at random with replacement from the selected deck and observes its color. This is done via an animation that shows the card leave and return to the deck.

2. Prediction: After observing that round's card, the participant reports their posterior belief that each deck was selected. The interface displays the exogenous information they have received: the wheel representing the prior, the history of observed cards in the task, and the decks. Previous reports in the current task or in completed tasks are not displayed.

[1] Experimental instructions are available in the Online Appendix.

Predictions can be any integer between 0 and 100 and must add to 100. The interface tracks the remaining percentage points to allocate as the participant updates their predictions. They must enter a prediction for both states — that is, the field for one state does not automatically populate based on the prediction for the other state.

2.1.1 Sequences

A participant is randomly assigned to one of two sets of sequences, where each set has its own task order. I determine 12 sequences in advance (6 for each set) based on the possible sequence compositions. Using 0 to denote an orange card and 1 to denote a blue card, these are:

0,0,0,0	0,1,1,1	0,0,1,1	0,1,1,1	1,1,1,1

The 12 slots are allocated proportional to each composition's expected frequency under p and q. Then, they are divided between sets to preserve this distribution as closely as possible. For the uniform prior, since signals are symmetrically informative:

$$p(\text{Deck } A \mid n_1 \text{ blue cards}, n_0 \text{ orange cards}) = 1 - p(\text{Deck } A \mid n_0 \text{ blue cards}, n_1 \text{ orange cards})$$

I therefore invert reports after majority 0 sequences to obtain their equivalent majority 1 reports.

Tables 2A and 2C list the sequences for each uniform prior treatment in terms of their majority 1 normalization.[2] The main design feature is that sequences present equivalent information in different orders. At $t = 3$ and $t = 4$, each participant sees three sequences that permute the position of the conflicting signal (the 0): at $t = 3$, these are $\{0,1,1\}$, $\{1,0,1\}$, and $\{1,1,0\}$; at $t = 4$, these are $\{1,0,1,1\}$, $\{1,1,0,1\}$, and $\{1,1,1,0\}$.[3]

Tables 2B and 2D list the sequences for the non-uniform prior treatments. The construction of these sequences exploits the relationship between the uniform and non-uniform problems. Specifically, a non-uniform prior $p = q$ is equal to the Bayesian posterior after observing one signal of precision q under a uniform prior.[4]

Accordingly, for each sequence in the uniform treatment with precision q, we create an informationally equivalent non-uniform sequence by removing the leading 1 signal and shifting the remaining three signals to the start. Tables 2A and 2B illustrate this for $q = .6$: the shaded cells in each row of Table 2B are the shift of the corresponding row in

[2] Appendix table A25 lists all sequences used in the experiment.

[3] It was not possible to include the ordered sequence $0,1,1,1$, given the number of available slots.

[4] To see this: $p(\text{Deck } A \mid blue) = \frac{p(blue \mid \text{Deck } A)\, p(\text{Deck } A)}{p(blue \mid \text{Deck } A)\, p(\text{Deck } A) + p(blue \mid \text{Deck } B)\, p(\text{Deck } B)} = \frac{q \cdot 1/2}{q \cdot 1/2 + (1-q) \cdot 1/2} = q$

Table 2A. Tables 2C and 2D show the same process for $q = .8$. This approach yields two representations of the same objective information at each t: a uniform prior with t signals, or a non-uniform prior (corresponding to one signal) with $t-1$ additional signals.[5]

2.2 One-shot inference tasks

After the sequential section of the study, the participant completes 11 one-shot inference tasks, where the first task is unincentivized practice. There are two differences between the one-shot and sequential section.

1. Task structure: The participant makes a prediction based on *one signal only*. Predictions are made via strategy method: prior to observing the drawn card, the participant reports a posterior conditional on the event that the drawn card is blue, and a posterior conditional on the event that the drawn card is orange. As noted earlier, q is fixed for all sequential and one-shot tasks.

2. Prior variation: Unlike in the sequential section, the exogenous prior changes each task. The participant completes one incentivized task with each of the following priors, excluding p in the sequential section:

$$p \in \left\{.5, .55, .6, .65, \tfrac{2}{3}, .7, .75, .8, .85, .9, .95\right\}$$

   We exclude the sequential p, since we already observe a one-shot update from that prior in round 1 of each sequence.

A participant's previous reports in the sequential section of the study, or in completed one-shot tasks are not displayed. As in the sequential section, participants are randomly assigned to one of two task orders.

2.3 Additional Details

Overall sequencing of study There are several short exercises after both inference sections: two risk aversion elicitations, free response questions about the decision-making process, and demographic questions.

At the start of the study, participants are told the study has two parts, both of which consist of prediction tasks. However, they do not receive a description or instructions for the one-shot section until they complete the sequential section. In the instructions for the

[5] Table 2D omits two sequences: $\{0,0,0,0\}$ and $\{0,0,1,0\}$. These provide no direct comparison to the corresponding uniform treatment, but were necessary to include to match the expected frequency of signals.

sequential section, participants must pass an attention check and a comprehension quiz within two attempts. Appendix figure A1 summarizes the timeline of the study.

Payment Predictions are incentivized, with payment calculated using the binarized scoring rule (Hossain & Okui, 2013). This payment rule is robust to risk aversion. Following Danz et al. (2022), participants are informed that this procedure guarantees they maximize their payment by reporting predictions truthfully. A detailed mathematical explanation of the rule is available via button in the instructions. For the one-shot task, the participant is paid for the guess that corresponds to the card that is actually drawn. That is, if a blue (orange) card is drawn, they are paid for their guess about the state conditional on a blue (orange) card.

Participants are paid for their decisions in half of the rounds of the sequential tasks, half of the one-shot tasks, and one of the two investment tasks. In each of the three cases, the paid tasks are chosen randomly. Participants receive a fixed participation payment of 4.5 USD, and any bonus payment from their choices that exceeds this amount.

2.4 Sample

Participants are recruited on the online platform Prolific. Each of the four treatments has 60 participants, split evenly across the two task orders. I pre-registered exclusion criteria based on mistakes for each section of the study. These are:

1. Updating in the wrong direction.
2. Not updating.
3. Reporting a posterior of q when it is not Bayesian to do so, which corresponds to complete base rate neglect.
4. When reporting beliefs by strategy method, entering the same beliefs for both signals.

Our main sample will be participants who make a mistake in less than half of their total decisions across the two sections. The final sample sizes are listed in Table 1.

3 Order independence in the aggregate

We begin with our test for order independence, which relies on variation in sequence order in the sequential inference task. We observe the following sets of signals in more than one order:

t	Sequence composition	Orders	Within subject
2	$\{0,1\}$	$\{0,1\}$ $\{1,0\}$	No
3	$\{0,1,1\}$	$\{0,1,1\}$ $\{1,0,1\}$ $\{1,1,0\}$	Yes
4	$\{0,1,1,1\}$	$\{1,0,1,1\}$ $\{1,1,0,1\}$ $\{1,1,1,0\}$	Yes

The important feature of these sequences is the within subject variation at $t = 3$ and $t = 4$. In both treatments, each subject observes the conflicting signal at all possible positions in sequences of length 3. In sequences of length 4, they observe the conflicting signal at $t \in \{2,3,4\}$. I will refer to the sequences $\{0,1,1\}$, $\{1,0,1\}$, and $\{1,1,0\}$ as the time 3 equivalent sequences, and the sequences $\{1,0,1,1\}$, $\{1,1,0,1\}$, and $\{1,1,1,0\}$ as the time 4 equivalent sequences.

Under order independence, individual i's time t posterior p_{it} is equal across all sequences with the same number of 0 and 1 signals. Concretely:

$$\begin{aligned}
\text{At } t=3: \quad & p_{i3} \,|\, \{0,1,1\} \quad = \quad p_{i3} \,|\, \{1,0,1\} \quad = \quad p_{i3} \,|\, \{1,1,0\} \\
\text{At } t=4: \quad & p_{i4} \,|\, \{1,0,1,1\} \quad = \quad p_{i4} \,|\, \{1,1,0,1\} \quad = \quad p_{i4} \,|\, \{1,1,1,0\}
\end{aligned}$$

We compare $\{0,1\}$ and $\{1,0\}$ between subjects, as a given individual does not see both orders.[6] The qualitative patterns are the same for both precisions, and the effects are not statistically distinguishable, so I will discuss the results in terms of the $q = 0.6$ treatment. The analogous figures and tables for $q = 0.8$ are in the Appendix.

Figure 3 shows a CDF of posterior reports after $\{0,1\}$ and $\{1,0\}$. Order matters substantially even after only two signals, consistent with other papers that examine short sequences (Agranov and Reshidi (2024); Chan (2025)). The mean posterior after $\{1,0\}$ is 8.2 pp lower than the mean report after $\{0,1\}$ ($p = 0.002$), and the distributions are significantly different (Kolmogorov-Smirnov $p = 0.029$).[7]

Moving to $t = 3$, we can begin to exploit our within subject variation. Figure 4 shows an individual level plot of reports for each of the time 3 equivalent sequences. Each 'column' of the plot corresponds to a participant, and the three markers correspond to

[6] The sequence $\{0,1,1,1\}$ is missing, and $\{0,1\}$ and $\{1,0\}$ are not observed within subject, due to the constraint of matching a signal's expected frequency. I opted to include $\{1,0,1,1\}$ over $\{0,1,1,1\}$ to observe consecutive positions for the conflicting information.

[7] For all analysis in the paper, corresponding regression tables are included in the Appendix.

the participant's reports in each of the sequences. Participants are sorted in increasing order of their report after $\{1,1,0\}$. The figure shows a striking pattern: while most people report a similar posterior after $\{0,1,1\}$ and $\{1,0,1\}$, a minority of the population makes a substantially lower report after $\{1,1,0\}$. The gap between $\{1,1,0\}$ and the other two sequences is 9 pp ($p = 0.000$). In contrast, the difference between $\{1,0,1\}$ and $\{0,1,1\}$ is only 0.2 pp.

At $t = 4$, we observe the exact same pattern. Figure A2 replicates Figure 4 with the three equivalent time 4 sequences, sorting on the report after $\{1,1,1,0\}$. Again, a minority of the population has an extreme response to the 0 signal. Reports after $\{1,1,1,0\}$ are about 10 pp lower than reports after $\{1,0,1,1\}$ and $\{1,1,0,1\}$ ($p = 0.011$ and $p = 0.001$, respectively), whereas the difference between $\{1,1,0,1\}$ and $\{1,0,1,1\}$ is only 1 pp. The latter result is notable given that just one period prior, there was a large discrepancy in beliefs between these two sequences — adding an additional 1 signal aligns reports again. Kieren et al. (2025) find a similar pattern of overresponse to conflicting information, immediately followed by 'correction', using between subject variation.

Overresponse to the conflicting signal is an individual-specific trait. Figure A3 shows a scatter of the within person order effect across time. The y-axis is the posterior after $\{1,1,1,0\}$ minus the averaged posterior after $\{1,0,1,1\}$ and $\{1,1,0,1\}$. The x-axis is the posterior after $\{1,1,0\}$ minus the averaged posterior after $\{0,1,1\}$ and $\{1,0,1\}$. There is a strong positive correlation — that is, those who exhibit order dependence at $t = 3$ are also likely to at $t = 4$ (correlation coefficient = 0.715 ($p = 0.000$)).

The individual level scatters additionally show that order dependence arises from a particular mistake. At $t = 3$ and $t = 4$, the signals favor state 1. When the sequence ends in a 0 signal, however, order dependent individuals frequently report a posterior for state 1 that lies below the initial prior of 0.5. In other words, they favor the state disfavored by the evidence. This means that violations of order dependence are closely tied to a second violation of Bayesian reasoning: holding a posterior on the wrong side of the prior.

Despite these deviations, participants in the $q = 0.6$ treatment are remarkably well calibrated on average. After the $t = 3$ equivalent sequences, the Bayesian posterior is 0.6; the average observed report is 0.58. After the $t = 4$ equivalent sequences, the Bayesian posterior is 0.69; the average report is 0.687. Even those who exhibit order dependence are within a few percentage points of the Bayesian posterior when the sequence does not end in 0. Together these results indicate that order dependence substantially but temporarily distorts beliefs for a minority of the population, and that this error may not necessarily reflect low quality updating more generally.

4 Prior sufficiency in the aggregate

In this section, we test for violations of prior sufficiency, continuing to aggregate the sample. To help explain the empirical strategy, let us return the manager example from the introduction. The top panel of Figure 2 shows the manager's inference problem in the case of the existing worker, Worker O (left), and the case of a newly arrived worker (right).

For the first period, the manager only observes s_1 from Worker O, updating his exogenous prior p^O to p_1^O. At $t = 2$, the first new worker $N1$ is hired. The manager's prior about $N1$ is exactly equal to p_1^O, and $N1$ and O generate the same signal s_2. Prior sufficiency implies that the posterior about $N1$, p^{N1}, should equal the posterior about O, p_2^O.

We can now repeat this logic for any period. Suppose at $t = 3$, a completely new worker $N2$ is hired, and the prior about $N2$ is equal to the current belief about O. Again, both workers send the same signal s_3. Prior sufficiency implies we should observe equivalent posteriors. The one-shot task allows us to construct this counterfactual exactly.

For each participant, we match each sequential decision p_{it}^{seq} to their one-shot task with prior equal to $p_{i,t-1}^{seq}$ and signal equal to s_t. We collect updates from a grid of priors that include all multiples of .05 between 0.5 and 1, as well as 2/3, and collect updates for both signals via strategy method. In practice, this means we match the vast majority of reports exactly. But, when no perfect prior match exists, we match to the closest prior on the grid.

Prior sufficiency implies:

$$p_{it}^{seq} \;\Big|\; \left\{s_t,\, p_{i,t-1}^{seq}\right\} = p_i^{one} \;\Big|\; \left\{s \,=\, s_t,\, p \,=\, p_{i,t-1}^{seq}\right\} \tag{1}$$

This test does not hinge on any assumption about the functional form of an individual's updating rule. An individual can use any procedure, including one that violates order independence, as long as this procedure results in the same posterior belief within the two tasks.

For the analysis, I will use the term "sequential posterior" to refer to p_{it}^{seq}. I will use the term "one-shot posterior" to refer to p_i^{one}. So, for example, the one-shot posterior for $\{1, 0, 1\}$ refers to the update in the one-shot task that is informationally equivalent to the sequential problem at $\{1, 0, 1\}$.

Given the results with respect to order independence, we will look separately at sequences ending in a 1 signal and sequences ending in a 0 signal. Figures A4 and 5 are individual level scatters of reports for the time 3 sequences $\{0, 1, 1\}$ and $\{1, 0, 1\}$, and the

time 4 sequences $\{1,0,1,1\}$ and $\{1,1,0,1\}$, respectively. Each 'column' is a subject and each marker corresponds to the person's average report in the sequential task and one-shot task after the sequences. The pairs of markers are ordered by the sequential average.

At $t = 3$, the average report in the sequential task is 3.8 pp higher ($p = 0.000$) than the average report in the one-shot task. This increases to 6.9 pp ($p = 0.000$) for $\{1,0,1,1\}$ and $\{1,1,0,1\}$. In contrast, there is no difference between tasks for $\{1,1,0\}$ and $\{1,1,1,0\}$. The sequential task is 1.7 pp higher and 3.4 pp higher, respectively, but neither of these are significant. We will see in the next section that this is because of heterogeneity.

The $q = 0.8$ treatment exhibits the same qualitative patterns as the $q = 0.6$ treatment, with larger magnitudes. For the sequences ending in 1, the sequential reports are 4.1 pp higher in $t = 3$ ($p = 0.009$), and 13.6 pp higher in $t = 4$ ($p = 0.000$). The sequential reports are also significantly more extreme than the one-shot reports for the sequences ending in zero: 7.5 pp higher ($p = 0.077$) and 16.4 pp higher ($p = 0.002$) for $\{1,1,0\}$ and $\{1,1,1,0\}$, respectively.

If we restrict the sample to sequential reports whose priors have an exact match in a one-shot task, the main results continue to hold. This removes 22 out of 705 decisions in the $q = 0.6$ treatment (3%) and 96 out of 720 in the $q = 0.8$ treatment (13%). The most common reason for exclusion is that the prior equals 1 (36% of excluded cases). Such cases occur more often under the higher precision. The only result that is sensitive to this restriction is in the $q = 0.8$ treatment: the gap between the sequential and one-shot tasks for $\{1,1,0\}$ and $\{1,1,1,0\}$ becomes less significant (for $\{1,1,0\}$, it becomes $p = 0.106$; for $\{1,1,1,0\}$, it becomes $p = 0.078$). This is unsurprising given that the observations that are being dropped are largely at the boundary. If we exclude priors at the boundary only, both remain signifinicant at the 10% level.

Together, these results indicate that prior sufficiency fails in a majority of the population, rather than a subset. Unlike order dependence, violations occur in sequences that end in a non-conflicting signal. Moreover, the magnitude of violations is fairly similar to the average order effect. In $q = 0.6$, the order effect is 9-10pp, while the average gap between the sequential and oneshot tasks is 6.9pp; in $q = 0.8$, the order effect is 14-15pp, which is on par with the $t = 4$ difference between the sequential and oneshot tasks.

5 Heterogeneity in violations

We will now examine the relationship between violations of prior sufficiency and violations of order dependence. Given clear heterogeneity in the population on the latter, I estimate a mixture model to partition individuals in a more principled way.

The model is non-parametric to remain agnostic about the rule that generates beliefs: I simply use a regression of the observed posterior on a vector of constants for each unique sequence at each time t.

$$\sum_j \beta_j 1[sequence = j] + \epsilon_{ij}$$

Note that this effectively reduces to using k-means, where the data for a given individual is just the vector of their posterior reports for each sequence. Below, I present results for two clusters. Including a third makes one cluster very small and adds no qualitative explanatory power.

5.1 Heterogeneity in order independence

The mixture model divides the population on the behavior of overreacting to a 0 signal. Figure 7 reproduces the individual level scatter of reports after the equivalent $t = 4$ sequences, now color coding by type. The model cleanly splits on the value of the report after $\{1, 1, 1, 0\}$.

Those in the minority cluster (14 of 47 subjects (30%)) exhibit strong order dependence: their average report after $\{1, 1, 1, 0\}$ is 38.1 pp lower than the average report in $\{1, 0, 1, 1\}$ and $\{1, 1, 0, 1\}$. Those in the majority cluster (33 of 47 subjects (70%)) show no order dependence. The individual level scatter for $t = 3$ (see Figure A5) shows the same pattern — while not as clean as the $t = 4$ figure, those classified into the minority cluster are much more likely to show a gap between $\{1, 1, 0\}$ and the other two sequences.

Figure A6 is the equivalent $t = 4$ scatter for the $q = 0.8$ treatment. The mixture model again partitions the population on overreaction to a 0 signal, in similar proportions. 37 of 48 subjects (77%) are in the majority cluster and the remaining 11 (23%) are in the minority cluster.

5.2 Heterogeneity in prior sufficiency

We will now test prior sufficiency separately for each type. There are two reasons to do this. First, we want to identify whether failures of prior sufficiency persist among otherwise "good" updaters. The majority type at least adheres to one principle of Bayesian reasoning — order independence — so if they deviate from prior sufficiency, it indicates that this type of mistake is widespread.

Second is that the experimental design allows us to ask whether, among those who exhibit order dependence, prior sufficiency is the source. This is a result that would

follow directly from a standard Grether (1980) type model, but it has not been considered in the literature. To understand how prior sufficiency could generate order dependence, let us consider our manager example again. Suppose there are two new workers, $N1$ and $N2$, and the manager has the same prior about both. $N1$'s first signal is H and $N2$'s first signal is L. After this signal, the manager should have a different belief about each worker because the signals are different. Now suppose that $N1$'s second signal is L and $N2$'s second signal is H. If the manager adheres to prior sufficiency but updates in a way that depends on the prior value (which, for example, Agranov and Reshidi (2024) suggests), then the posteriors may remain different after the second signal, even though the aggregate information received about each worker is the same.

We will begin with the $q = 0.6$ treatment and majority type. Figures 8-10 plot the time series of mean reports in the sequential task and one-shot task for the sequences $\{1,0,1,1\}$, $\{1,1,0,1\}$, and $\{1,1,1,0\}$, respectively. A gap emerges even after just two signals: the mean sequential report for $\{1,1\}$ is 3.1 pp higher ($p = 0.020$).

By the third signal, we observe gaps in all possible sequences. In all cases, the sequential report is more extreme than the one-shot report. This ranges from 2.7 pp ($p = 0.034$) for $\{0,1,1\}$ to 8.7 pp ($p = 0.000$) for $\{1,1,0\}$.

The gap persists from $t = 3$ to $t = 4$ in each of the time 4 equivalent sequences. In the case of $\{1,0,1,1\}$ and $\{1,1,0,1\}$, the final sequential report is 4.2 pp ($p = 0.058$) higher and 10 pp ($p = 0.000$) higher, respectively. For $\{1,1,1,0\}$, the final sequential report is 13.1 pp ($p = 0.000$) higher.

The discrepancy between the tasks holds across the entire distribution. Figures A7-A8 and 11-12 plot CDFs for the two tasks for each of the $t = 3$ and $t = 4$ equivalent sequences (pooling the sequences ending in a 1 signal). In each case, the distribution of sequential reports first order stochastic dominates the distribution of one-shot reports. CDFs of the within person difference in reports confirm that this gap arises from the majority of subjects making a higher sequential report (see Figures A9 and A10 for the $t = 3$ and $t = 4$ equivalent sequences, respectively).

For the minority type, our smaller sample size limits statistical power. However, two key patterns at $t = 4$ demonstrate that these subjects' sequential behavior is inconsistent with prior sufficiency. Figure 13 shows CDFs of the sequential and one-shot reports after $\{1,1,0,1\}$. The distribution of sequential reports first order stochastic dominates, with sequential reports exceeding one-shot reports by 13.4 pp on average ($p = 0.001$). This indicates that the immediate correction following a downward update occurs specifically within the sequential context. When the identical information arrives in a one-shot format, beliefs increase far less.

A complementary pattern emerges for $\{1,1,1,0\}$. Figure 14 compares the sequential and one-shot CDFs for this sequence. Here we observe the reverse ordering: the distribution of one-shot reports first order stochastic dominates, with a mean that is 19.4 pp higher ($p = 0.000$). The minority cluster's other characteristic behavior — overresponding to conflicting signals — thus also manifests only in sequential settings. The corresponding one-shot problem does not induce nearly as large a downward update.

The $q = 0.8$ treatment replicates these qualitative patterns for both the majority and minority types. For the majority type, sequential reports are more extreme across all sequences. And for the minority type, sequential reports substantially exceed one-shot reports for $\{1,1,0,1\}$, while the reverse holds for $\{1,1,1,0\}$.

To summarize, several patterns hold consistently across signal precisions. The mixture model partitions the population precisely on adherence to order independence. Both types violate prior sufficiency, though in distinct ways. The majority type — those satisfying order independence — report systematically higher posteriors in the sequential task, regardless of information ordering. The minority, in contrast, show inverse behavior relative to the sequence: their report is higher in the sequence when they are compensating for a downward update and lower in the sequence when they are observing a conflicting signal. Finally, these effects do not improve, and sometimes become worse, as the sequence gets longer.

6 Prior sufficiency with respect to a sequence

The one-shot task only permits a test of whether subjects adhere to prior sufficiency period by period. We now exploit variation in the initial prior to test prior sufficiency with respect to a *sequence*, rather than a single signal. In this case, our counterfactual problem is also a sequential problem, making it more similar in format to the sequential task in the uniform treatment.

6.1 Empirical strategy

This test will compare reports in the uniform treatment with precision q to reports in the non-uniform treatment with precision q.

First, I will explain what counterfactual we are aiming to test, using our manager example one more time. Figure 2 reproduces our baseline sequential problem on the left, and the new sequential counterfactual on the right. Suppose that at $t = 2$, a new worker N is hired. The prior about N, p_N, is equal to the current belief about O, p_1^O. Between

$t = 2$ and $t = t'$, the manager receives the same sequence of signals about the two workers. If the manager adheres to prior sufficiency, it should be the case that the $t = t'$ posterior about each worker is the same. This is because there exists some time at which the prior value was the same (in our example $t = 2$), and then all of the intervening signals are the same.

We will use the non-uniform prior treatments to generate this counterfactual between person. Namely, for each sequence in the uniform treatment, the non-uniform treatment has an informationally equivalent problem with prior of q. I construct these problems by dropping the first signal from the corresponding uniform sequence and shifting the remaining three signals from times 2–4 up to times 1–3. For example, $\{1, 0, 1, 1\}$ in the uniform treatment becomes $\{0, 1, 1\}$ in the non-uniform treatment, where the prior of q substitutes for the dropped first signal. These problems are equivalent because a prior of q is just the Bayesian posterior after 1 signal. All that varies is whether the first piece of information comes in the form of a signal (uniform treatment), or whether it is implied in the prior (non-uniform treatment).

Given this process, we have the follow paired problems for each precision:

(a) $q = 0.6$

Uniform seq.	**Non-uniform seq.**
{1,0,1,1}	{0,1,1}
{1,1,0,1}	{1,0,1}
{1,1,1,0}	{1,1,0}
{0,1,1,0}	{1,1,0}
{1,1,0,0}	{1,0,0}

(b) $q = 0.8$

Uniform seq.	**Non-uniform seq.**
{1,0,1,1}	{0,1,1}
{1,1,0,1}	{1,0,1}
{1,1,1,0}	{1,1,0}
{0,1,1,0}	{1,1,0}
{1,1,1,1}	{1,1,1}

I will use the convention of referring to time periods and sequences from the perspective of a subject in the uniform treatment. For example, a report after $\{1, 0, 1, 1\}$ will refer to the uniform subject's report after observing all of $\{1, 0, 1, 1\}$, and will refer to the non-uniform subject's report after observing the informationally equivalent sequence $\{0, 1, 1\}$. Likewise, $t = 4$ would refer to the uniform subject's report at time 4, and the non-uniform subject's report at time 3.

We can perform two different tests comparing the uniform and non-uniform distribution of reports at each t for a given sequence. The first is an exact test. If a subject in the uniform treatment reports the Bayesian posterior after the 1st signal, his endogenous

prior at $t = 2$ is exactly equal to the non-uniform subject's exogenous prior. This tracks Figure 2 exactly, save that the comparison is between subject rather than within subject. 60% of initial reports in the $q = 0.6$ treatment are the Bayesian posterior, making this test viable for that treatment.

The second is a test based on the ordering of the distribution of reports from each treatment, which we can apply for both precisions. It is based off the following observation. In the $q = 0.8$ uniform treatment, the mean report after 1 signal is about 70%, which is 10 pp lower than the initial prior in the corresponding non-uniform treatment. Only 30% of reports are equal to 0.8 exactly, and 97% of reports are less than or equal to 0.8.

Thus, for any individuals U and N randomly chosen from the uniform and non-uniform treatments, respectively, N's prior p_N (exogenously set to 0.8) will be weakly higher than U's prior p_U. In the two state setting, we can rewrite this as:

$$\frac{p_N(\omega = 1)}{p_N(\omega = 0)} \geq \frac{p_U(\omega = 1)}{p_U(\omega = 0)} \tag{2}$$

Given that reports have this property, we can test for prior sufficiency under a restricted class of updating rules, using the following result from (Chan, 2025): If p_N and p_U are two prior distributions that satisfy (2), then if the updating rule has the form

$$p'(\omega_j \mid s_t) = \frac{\left[p(\omega_j)\right]^{\beta} G_j(s_t)}{\sum_k \left[p(\omega_k)\right]^{\beta} G_k(s_t)} \quad \text{where } G_k(\cdot) > 0 \text{ for all } k \tag{3}$$

the corresponding posteriors p'_N, p'_U satisfy

$$\frac{p'_N(\omega = 1)}{p'_U(\omega = 0)} \geq \frac{p'_N(\omega = 1)}{p'_U(\omega = 0)}$$

Note that (3) is a generalized form of Grether (1980), where the signals can be distorted according to any positive function $G_k(\cdot)$. In our two state setting, this result is equivalent to saying: if N's prior is more extreme than U's and updating follows (3), then N's posterior is also more extreme than U's. We have already established that for any pairing of individuals from the two treatments, N's prior is more extreme. So, if people use a rule with the form of (3), the distribution of non-uniform reports should always remain above the distribution of uniform reports. This should hold even if each i has their own β_i and own $G_{k,i}(\cdot)$ function. Due to randomization, these unobserved individual level parameters are, in expectation, balanced across the two treatments. If the posterior distributions do not have the same ordering as the prior distributions, updating

is inconsistent with being prior sufficient under a rule with the form of (3).

6.2 Results

As with the one-shot task, we find deviations from prior sufficiency. We will first show results for the exact test of prior sufficiency, comparing reports in the $q = 0.6$ treatment from sequences where the first report is the Bayesian posterior to reports in the non-uniform for the corresponding sequence.

Exact test Figures 15-17 and A11 plot the time series of the average sequential report in the two $q = 0.6$ treatments. For reference, these figures also include the average one-shot report in the uniform treatment, restricting to those sequences where the first report is the Bayesian posterior.

The main patterns distinguishing the sequential and one-shot tasks also emerge in the comparison of the uniform and non-uniform. Namely, as time passes, discrepancies emerge in new sequences, and once a discrepancy on a given path exists, it almost never improves as long as the sequence remains informative.

By $t = 4$, there are gaps for all three of the equivalent sequences $\{1,0,1,1\}$, $\{1,1,0,1\}$, and $\{1,1,1,0\}$. The difference between the non-uniform and uniform treatment is 0.055 ($p = 0.003$), 0.043 ($p = 0.059$), 0.093 ($p = 0.028$), respectively. CDFs show that in these sequences, the distribution of uniform reports first order stochastic dominates the distribution of non-uniform reports, as in the one-shot comparison (see Figures A12, A13, and A14).

When the sequence becomes uninformative, existing discrepancies resolve and new ones do not appear. For $q = 0.6$, we have two uninformative sequences: $\{0,1,1,0\}$ and $\{1,1,0,0\}$. In the case of $\{0,1,1,0\}$ (in Figure A15), the belief in the non-uniform is initially more extreme, and in the case of $\{1,1,0,0\}$ (in Figure A16), the uniform report is significantly higher in both periods 2 and 3. However, by the end of each sequence, there is no gap.

Distribution based test for q=0.8 We now apply our distribution based test to the $q = 0.8$ treatments. In most cases, in $t = 2$ and $t = 3$, the distribution of reports in the non-uniform treatment remains to the right of the distribution of reports in the uniform treatment. This is in line with the fact that the non-uniform subjects start almost exclusively higher than any uniform subject.

By $t = 4$, however, the distribution for the uniform lies to right of the distribution for the non-uniform for both $\{1,0,1,1\}$ and $\{1,1,0,1\}$. Figure 18 shows the CDF for each treat-

ment, pooling these sequences. Reports in the uniform are 0.049 pp higher ($p = 0.019$), and the distributions are significantly different.[8] Figure 19 shows the corresponding CDFs for the sequence $\{1,1,1,0\}$, where the markers for the uniform treatment are color-coded by type. The mean difference between the two treatments is small, but the distributions are significantly different. We can immediately see that this arises from the divergence in behavior between the majority and minority clusters. The majority is above the uniform report, while the minority is below it.

Distribution based test for q=0.6 In the $q = 0.6$ treatment, 80% of time 1 posterior reports are no larger than the Bayesian posterior. While not as clean as the $q = 0.8$ case, we apply the distribution based test here as well.

The qualitative results from the exact test carry over for $\{1,0,1,1\}$ and $\{1,1,0,1\}$, although the magnitude of the gap between the uniform and non-uniform is smaller. Figure 20 shows the distributions in each treatment for the two sequences pooled: the mean difference is 0.039 ($p = 0.019$), and the distributions are significantly different from each other (K-S $p = 0$).

The gap between the uniform and non-uniform treatments for sequences ending in 0, however, is now insignificant. The figures from the $q = 0.8$ treatments immediately make clear why this is. When we restrict the sample, we incidentally exclude observations from minority type subjects who report a very low belief after a conflicting signal. Figures A19 and A20 reproduce the CDFs for $\{1,1,0\}$ and $\{1,1,1,0\}$, color-coding by type. We can see that the reports lying above the non-uniform CDF almost exclusively belong to the minority type. The distributions are not significantly different in the case of $\{1,1,1,0\}$, but they are at the 10% level in the case of $\{1,1,0\}$.

In summary, participants fail the sequence based test of prior sufficiency, and do so in the same direction as in the sequential vs. oneshot comparison: reports in the uniform treatment are higher than reports in the informationally equivalent non-uniform sequence. If we use the entire sample, without conditioning on type or time 1 posterior, the final report is between 4 and 5 pp higher in the uniform treatment for sequences ending in a 1 signal. If we restrict to reports from sequences where the time 1 report is the Bayesian posterior, the average report in the uniform treatment is between 4 and 9 pp higher. So, even when we make the format of the counterfactual problem much closer to the format of the sequential problem, we still observe that beliefs in the sequential problem are more extreme.

[8] Figures A17 and A18 are the CDFs for $\{1,0,1,1\}$ and $\{1,1,0,1\}$ separately. In both cases, the qualitative direction is the same. However, the mean difference for $\{1,0,1,1\}$ is not significant, only the distributions are significantly different.

7 Estimation of updating rules

The benchmark empirical model used in the updating literature, Grether (1980), implies that dynamic updating follows an AR1 structure — namely, the posterior is written as a function of the prior and the most the recent information. This functional form is only correctly specified if prior sufficiency holds.

Our results suggest that such a model will fail to organize the data well. In this section, we will estimate the Grether model under a variety of parameterizations, and see that, consistent with our non-parametric findings, it does a poor job. The models match the mode of the distribution, but underestimate the frequency of the modal value. Improvements on matching the mode tend to come at the cost of matching the rest of the distribution less accurately. The first error is particularly important because for the majority type, the modal value is often very close to Bayesian — meaning that this parametric family of models overpredicts the frequency of mistakes.

7.1 Extending Grether to sequential inference

Let us start with the static version of Grether. The model assumes that an individual follows Bayes' rule, with some power distortion on the prior and signal. Assuming two states, with prior p, the posterior p' is given by:

$$p'(\omega|s) = \frac{p(s_|\omega = 1)^\beta \; p(\omega)^\delta}{\sum_{\omega' \in \{0,1\}} p(s|\omega')^\beta \; p(\omega')^\delta}$$

We then divide the posterior for state 1 by the posterior for state 0 and take the log. This gives us an expression in log odds:

$$\overbrace{\log\left(\frac{p(\omega = 1|s_i)}{p(\omega = 0|s_i)}\right)}^{\pi'} = \beta \overbrace{\log\left(\frac{p(s|\omega = 1)}{p(s|\omega = 0)}\right)}^{\lambda} + \delta \overbrace{\log\left(\frac{p(\omega = 1)}{p(\omega = 0)}\right)}^{\pi}$$

To economize on notation, I will use λ to refer to the log-likelihood ratio, and π to refer to the log odds of the states.

Now, let us extend the model to multiple periods. For simplicity, we will assume a uniform prior, so that the prior term drops out. The model is an AR1 time series process: the prior on the right hand side is the first lag of the outcome. We can observe the key

implication of this structure by recursively substituting in for the prior:

$$
\begin{aligned}
&t = 1 : \pi_1 = \beta\lambda_1 \\
&t = 2 : \pi_2 = \beta\lambda_2 + \delta\pi_1 = \beta\lambda_2 + \delta\Big(\underbrace{\beta\lambda_1}_{\pi_1}\Big) = \beta\lambda_2 + \delta\beta\lambda_1 \\
&t = 3 : \pi_3 = \beta\lambda_3 + \delta\pi_2 = \beta\lambda_3 + \delta\Big(\underbrace{\beta\lambda_2 + \delta\beta\lambda_1}_{\pi_2}\Big) = \beta\lambda_3 + \delta\beta\lambda_2 + \delta^2\beta\lambda_1 \\
&\qquad \vdots \\
&t = t' : \pi_{t'} = \beta\lambda_{t'} + \sum_{j=1}^{t'-1} \delta^{t'-1-j}\beta\lambda_j
\end{aligned} \tag{4}
$$

What equation (4) shows is that under the model, the posterior belief is simply a polynomial of the weights β and δ and the signals.

Both the sequential vs. oneshot comparison and the uniform vs. non-uniform comparison test for prior sufficiency under a more general model than (4) — where the oneshot test is always completely non-parametric. The fact that we find violations of prior sufficiency in both cases indicates that any Grether style linear model is misspecified. This would be the case even with individual heterogeneity in the weights β and δ — the oneshot test is already at the individual level, and the non-uniform test is robust to heterogeneity.

We can assess the performance of Grether, as it turns into a linear econometric model simply by assuming an additive error term ϵ_{it}. As an econometric model, Grether is a dynamic panel, which creates an endogeneity challenge. I will address this in the estimation using a structural approach. However, I first will explain the problem more precisely.

Suppose the true model has a linear structure but with individual level heterogeneity. In other words, we believe that each individual's report is generated by a rule $\pi_{it} = \beta_i\lambda_{it} + \delta_i\pi_{i,t-1} + \epsilon_{it}$, where $\{\beta_i, \delta_i\}$ are individual weights and ϵ_{it} is classical measurement error. In this case, if we estimate OLS pooling all observations, the error term has the following structure

$$
\epsilon_{it}^{pooled\ OLS} = (\beta_i - \beta)\lambda_{it} + (\delta_i - \delta)\pi_{i,t-1} \qquad \text{where } \beta \text{ and } \delta \text{ are the estimated parameters.}
$$

This error term will be correlated with the lagged posterior, as they both contain an individual's weights (this is clear for the lagged outcome after recursive substitution, as in (4)).

If the weights β_i, δ_i are correlated, then both $\beta^{pooled\ OLS}$ and $\delta^{pooled\ OLS}$ are not consistent estimates of the means $E[\beta_i]$ and $E[\delta_i]$. Existing papers take the strategy of instrumenting $\pi_{i,t-1}$ with the Bayesian prior. This, however, does not recover the means of the weights either.[9] The econometric literature on random coefficient panel data models has recognized for some time that in these contexts no valid instrument exists (Hsiao & Pesaran, 2004). The only paper that proposes a non-parametric strategy for estimating moments of the distribution of random coefficients is (Lee, 2025), which recovers an identified set.

7.2 Econometric strategy

I will take a structural approach to estimating Grether, using simulated maximum likelihood. The reason for a structural approach is to simulate data under the model to assess its performance.[10] To illustrate the procedure, let us consider the simplest possible parameterization: each individual i is characterized by a parameter vector $\theta_i = \{\beta_i, \delta_i\}$, where β_i is a weight on the signal and δ_i is a weight on the prior.

I assume that the parameters are distributed multivariate normal, and that the additive error term ϵ_{it} is normally distributed with mean 0.

$$\begin{pmatrix} \beta_i \\ \delta_i \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_{\beta_i} \\ \mu_{\delta_i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\beta_i} & \sigma_{\beta_i\delta_i} \\ \sigma_{\beta_i\delta_i} & \sigma^2_{\delta_i} \end{pmatrix} \right) \tag{5}$$

$$\epsilon_{it} \sim_{iid} N(0, \sigma^2)$$

Using ϕ to denote the normal pdf, the log likelihood of an observation is:

$$\mathcal{L}_{it} = \ln(\phi\left[\pi_{it} - \theta_i x_{it}\right]) \qquad \text{where } x_{it} = [\lambda_{it}, \pi_{i,t-1}(\theta_i)]'.$$

We do not observe θ_i, however, so we cannot calculate the likelihood of a given observation directly. We need to integrate θ_i out so that the likelihood is only written in terms of

[9] See Appendix C.1 for details on endogeneity and IV estimation.

[10] Bland and Rosokha (2025) also adopt a structural approach to estimating moments of the distribution of weights, but use a Bayesian framework.

observed variables:

$$\mathcal{L}_{it} = \ln\left(\int \phi\left[\pi_{it} - \theta_i x_{it}\right] dF(\theta_i)\right)$$

This integral does not have an analytic solution, so we simulate it via numerical integration. We draw a grid of θ values, with each draw denoted s. We calculate the likelihood of a given observation (π_{it}, x_{it}) at each draw θ_s and then average. This gives us a simulated log-likelihood for each it:

$$\mathcal{L}_{it}^{sim} = \ln\left[\frac{1}{S}\sum_{s} \phi\left[\pi_{it} - \theta_s x_{it}\right]\right]$$

The estimation routine solves for the parameter vector that maximizes the simulated log-likelihood $\sum_i \sum_t \mathcal{L}_{it}^{sim}$. That is, the output is an estimate of the moments of the distribution of coefficients, as well as the variance of the error term.

I estimate the two parameter baseline model given in (5), as well as a three parameter baseline model motivated by the results so far. This model has a weight for conflicting signals, a weight for non-conflicting signals, and a prior weight. I use two different definitions of conflicting signals, and estimate a number of variants of each model, permitting time specific changes in the mean or variance of the individual parameters.[11]

The estimation follows the Grether model literally, using the recursive substitution representation in (5). So, for a given observation, the residual will in fact be written as:

$$\pi_{t'} - \left(\beta\lambda_{t'} + \sum_{j=1}^{t'-1} \delta^{t'-1-j}\beta\lambda_j + \pi\right) \qquad \text{where } \pi \text{ is the initial exogenous prior log odds.}$$

That is, I never subtract the observed lagged posterior. Beliefs only appear as the outcome. Additionally, I will assume that ϵ_{it} is classical measurement error — this means that lagged error terms will not appear in the likelihood. If they did, we would have to integrate them out as well, as they are unobserved.

I estimate each model on all participants and on the majority type only. For a given model, we simulate 1000 individuals per actual participant in the sample, using the functional form in 4, evaluated at an observed sequence, as well as a $\{\beta_i, \delta_i\}$ and ϵ_{it} drawn from the solution distributions.

[11] The full list of models can be found in Appendix C.2.

7.3 Results

Given that the simulated data is highly continuous and the observed data is not, I bin observations by 5 percentage point bins. I will use two different measures of fit to organize the discussion. Denote the fraction of observed data in bin j by $\theta^{obs.}_{bin\,j}$ and the fraction of simulated data in bin j by $\theta^{sim.}_{bin\,j}$ The first measure, which I will refer to as the mode gap, is: $\left(\theta^{obs.}_{bin\,j} - \theta^{sim.}_{bin\,j}\right)$ for the modal bin in the observed data. This is just the amount that the model over or underpredicts the weight on the mode. The second measure, which I will refer to as the non-mode gap is: the sum of $abs\left(\theta^{obs.}_{bin\,j} - \theta^{sim.}_{bin\,j}\right)$ over all non-modal bins, divided by 2. This measures the mass in the simulated distribution that would have to be reallocated to match the observed distribution outside of the modal bin. The division by 2 is simply a normalization, as the raw sum double counts mass.

Let us start just by considering a single signal. Figure 21a shows the observed and simulated distributions in the full sample for the baseline model with 1 signal weight. The gray bars show the frequency of observed data in each bin, and the scatter shows the fraction of simulated data in the bin. On the simplest sequence, the model has very poor fit: it underestimates the mode by about 30 pp and overestimates the mass in bins around the mode, with the non-mode gap equal to 24 percentage points.

A broader look across all sequences confirms that underestimation of the mode and misestimation of the distribution is widespread. Column 1 of Table 3 lists the mode gap for each sequence and Column 1 of Table 4 lists the non-mode gap for each sequence. In 11 of 12 sequences, the mode is underestimated by at least 10 pp. The non-mode gap is between 16.5 and 36.5 pp, with half of the sequences having a gap above 25 pp.

The longer informative sequences tend to have a smaller mode gap, but not necessarily a smaller non-mode gap. In other words, improvement on one dimension of fit does not translate to improvement on the other. Figures 21a-b compare $\{1\}$ to $\{1,1,0,1\}$ under the baseline model with 1 weight to illustrate. The sequence $\{1,1,0,1\}$ has a much smaller mode gap of 4.3 pp, but for both sequences, the non-mode gap is $\approx$25 pp.

Varying the parameterization of the model does not lead to substantial improvement. Given that we know there is extreme overresponse to a conflicting signal in the full population, one might expect that enriching the baseline model with a second signal weight for 'conflicting' signals would help. It does improve fit, but only marginally. The left panel of Table 5 shows the difference between the 2 weight and 1 weight model for our two fit measures.[12] The 2 weight model improves the non-mode gap for the short sequences of

[12] A signal is coded as conflicting if it requires an update in the opposite direction of the previous update. The Online Appendix includes alternative definitions of a conflicting signal.

length 1 and 2 and the sequences of length 4. It also mostly improves the mode gap, and for the sequences where it does not, it adds no more than 2.9 pp. Despite this, the same qualitative patterns persist under the 2 weight model: longer informative sequences still have smaller mode gaps but not substantially smaller non-mode gaps. This persistence is evident in Figures 21c-d, which compare $s = \{1\}$ and $s = \{1, 1, 0, 1\}$ for the 2 weight model.

A natural hypothesis for the poor fit would be that we are pooling two very different kinds of updaters when using the full sample. However, when we estimate the model only on the majority type, we observe similar results. In the simplest sequence of 1 signal, the modal bin is underestimated by 40 pp. Considering all of the data, longer informative sequences have a smaller mode gap, without clear improvement on non-mode fit. Figures 22a-b reproduce the comparison of $\{1\}$ and $\{1, 1, 0, 1\}$ under the 1 weight baseline model, estimated only on the majority type.

Furthermore, adding a second signal weight in the majority type sample produces much more mixed results than in the full sample. The right panel of Table 5 shows that this additional flexibility substantially worsens the non-mode gap for sequences of length 4 and the mode gap for certain sequences, suggesting limited benefits from more parameters. Models with time varying parameters, included in the Online Appendix, do not systematically resolve these fit issues.

The following exercise makes stark the limitation of the Grether model. If we estimate the baseline Grether specification for the full sample on the first signal and first update only, the predicted report falls in the $[0.6, 0.65)$ bin, correctly identifying the mode. Now suppose we assumed each simulated individual reports exactly this predicted value. This degenerate distribution would actually match the observed distribution better than the model estimated on all the data. The one period Grether simulation would misallocate 40 pp — precisely the amount of observed mass that is not on the mode. The full simulation, by contrast, is off by 30 pp on the mode and then overpredicts the rest of the bins by about 40 pp in total. The model can correctly identify where beliefs concentrate from one period alone, but adding more data worsens rather than improves its predictions.

Together, these exercises demonstrate that the parametric family of Grether models struggles to fit the data under the restrictions imposed by an AR1 process. The model consistently underpredicts the concentration of probability mass at the mode while simultaneously failing to capture the shape of the distribution around the mode. These failures persist across different parameterizations, different sample restrictions, and different sequences.

8 Sources of prior sufficiency violations

8.1 Aggregating signals

In this final section, we will consider some explanations for failures of prior sufficiency, specifically among the majority type. One candidate explanation is that individuals aggregate all signals received thus far with the initial prior, rather than updating from their current prior using only the most recent signal.

The main exercise makes use of variation in the timing of equivalent sets of signals. Recall that due to the structure of the problem, the Bayesian posterior depends only on the initial prior and the net count of 1 signals. For example, a sequence with two 1 signals and one 0 signal induces the same Bayesian posterior as a single 1 signal.

After cancellation, the sequences induce four unique Bayesian posteriors, where I refer to the reduced set of signals after cancellation as the "reduced" sequence. We observe the first three reduced sequences at an earlier period and at a later period.

Full sequence	Reduced sequence	Observed at
$\{0,1\}$ $\{1,0\}$ $\{1,0,0,1\}$ $\{1,1,0,0\}$	initial prior	$t = 2$ and $t = 4$
$\{0,1,1\}$ $\{1,0,1\}$ $\{1,1,0\}$	$\{1\}$	$t = 1$ and $t = 3$
$\{1,0,1,1\}$ $\{1,1,0,1\}$ $\{1,1,1,0\}$	$\{1,1\}$	$t = 2$ and $t = 4$
$\{1,1,1\}$	$\{1,1,1\}$	$t = 3$

Suppose that the report after a given reduced sequence depends on time, e.g., that the average report after $\{1,1\}$ is not equal to the average report after $\{1,1,0,1\}$. In this case, there are only two possible explanations: the prior is different (endogenously), and more signals have accumulated without changing the objective posterior. We can control for the former through the matched one-shot task. Any residual indicates that signal accumulation matters on its own.

First, we will establish that the reports do depend on time and do so in a consistent direction. Figure A21 plots CDFs of the initial report and the report after the $t = 3$ equivalent sequences for the majority type in the $q = 0.6$ uniform treatment. Here, we already see a small gap of 2.4 pp ($p = 0.010$). Figure 23 plots CDFs of the report after $\{1,1\}$

and the $t = 4$ equivalent sequences. The gap increases to 7.2 pp ($p = 0$), and the time 4 distribution is now essentially fully shifted to the right of the $t = 2$ distribution.

For the reduced sequence $\{1, 1\}$, we have a one-shot update at both points in time. So, we can compare the difference between the sequential and oneshot report at each period. If prior sufficiency explains the gap between the earlier and later period, the difference-in-difference estimate should be zero.

Figure 24 performs this comparison. The gap between the sequential and oneshot tasks is higher at time 4, with a residual of between 4 and 10 pp depending on the sequence. Indeed, Figure A22 replicates Figure 23 with one-shot reports for the sequences with two 1 signals — unlike for the sequential reports, the average one-shot report does not depend on time.

Together, these results indicate that the increase in reports over time cannot be accounted for with prior sufficiency. The mere fact that more signals have accumulated must matter in and of itself. The uninformative sequences (i.e., those with the same number of 0s and 1s) suggest a mechanism: that participants consider signals in the aggregate using a heuristic that considers the total count of signals.

Figure A23 shows CDFs of the reported belief after uninformative sequences at $t = 2$ and $t = 4$. In both cases, the overwhelming majority of participants report 0.5, which is the Bayesian posterior. So, when signals are exactly balanced, at least in sequences of this length, people cancel signals fully. However, when the sequence is informative, the total count of signals, rather than just the net count of 1 signals, seems to matter. This suggests that participants may employ a frequency-based heuristic, treating a higher count of total signals as providing additional information when those signals favor one state.

8.2 Updating in the non-uniform treatments

The results in the previous subsection suggest that individuals aggregate signals in some form. This generates a puzzle: if individuals aggregate signals in a sequence, then we might expect reports in the uniform and non-uniform treatments to converge over time as the accumulated signals come to dominate the initial prior. Instead, we observe the opposite. The gap between treatments grows larger as the sequence progresses. While I cannot fully reconcile these two facts, examining updating behavior in the non-uniform treatment provides suggestive evidence about what drives this divergence.

A possibility is that in the non-uniform treatment, individuals update in a manner much closer to the one-shot benchmark. However, for the sequences shared across the uniform and non-uniform treatments, we again find that reports in the sequential task

tend to be higher. Figures 25 and A24 plot CDFs of reports in the sequential and one-shot tasks for the pooled sequences $\{1,0,1,1\}$ and $\{1,1,0,1\}$ in the non-uniform $q = 0.6$ and $q = 0.8$ treatments, respectively. Reports in the sequential task are 5 pp higher for both precisions, and the sequential distribution first-order stochastically dominates. Figures A25 and A26 show the CDFs for $\{1,1,1,0\}$. There is no gap for $q = 0.6$, but sequential reports in the $q = 0.8$ treatment are 14 pp higher.[13]

We can also test whether signal accumulation matters independently of the prior by comparing the gap between sequential and one-shot reports across time for a given reduced sequence. For $q = 0.6$, we observe the reduced sequence $\{1\}$ and $\{1,1,1\}$ at $t = 3$ and $t = 5$.[14] For $\{1,1,1\}$, the sequential report exceeds the oneshot report by 3.5 pp at $t = 3$ ($p = 0.000$) and by 9 pp at $t = 5$ ($p = 0.000$), yielding a difference-in-differences estimate of 5.7 pp ($p = 0.000$). For $\{1\}$, the DID estimate is close to zero and insignificant.

Together, these results indicate that in both the uniform and non-uniform treatments, individuals do not update in a purely incremental fashion. This would mean that the exogenous parameters of the non-uniform treatment must lead to different updating behavior — we have one piece of evidence that this is the case.

In the non-uniform treatment, most subsequences contain signals that are either evenly split or favor the state that is ex-ante more likely. However, some subsequences contain a majority of 0 signals, contradicting the initial exogenous prior. For example, participant see $\{0,1\}$ and $\{1,0\}$, corresponding to a uniform prior with $\{1,0,1\}$ and $\{1,1,0\}$, while they see $\{0,0\}$, corresponding to a uniform prior with $\{1,0,0\}$. We can invert reports here to make the sequences comparable: the posterior for state 1 given $\{0,1,1\}$ (not directly observed) is equivalent to the posterior about state 0 given $\{1,0,0\}$ (directly observed). After inversion, order independence says the reports for these three sequences should be the same.

This equivalence does not hold in practice for informative sequences, generating apparent "order" effects that are different from the uniform treatments. Figure 26 shows CDFs of reports after $\{1,0,0\}$ (post-inversion) and $\{1,0,1\}$ for the $q = 0.6$ treatment, including all participants.[15] The distribution of reports for $\{1,0,0\}$ is shifted right (K-S $p = 0.05$), indicating more extreme beliefs when the sequence is presented as two 0 signals. This contrasts sharply with the uniform treatment, where the distributions are essentially identical (see Figure A27).

This pattern is even more pronounced for longer sequences. When comparing se-

[13] Figures A29-A32 show the equivalent CDFs for time 3. Reports are weakly more extreme in the sequential task in all cases.

[14] Note here that I am keeping the convention of counting time from the perspective of a uniform prior.

[15] Tables 6 and 7 list all equivalent sequences where we observe cases with majority 0s and majority 1s.

quences with three 0s and two 1s to sequences with three 1s and two 0s, reports are 18 pp higher for sequences with majority 0s (see Figure 27). Importantly, among sequences with composition $\{0,0,1,1,1\}$, there is no order effect — the gap emerges purely based on the number of 0 signals in the sequence.

The $q = 0.8$ treatment shows related patterns. Reports after $\{0,0,0,1,1\}$ are 20 pp higher than after $\{0,0,1,1,1\}$, where again there is no order effect for sequences with two 0 signals. Reports after $\{0,0,1\}$ are 6 pp lower than after $\{1,0,1\}$, opposite to $q = 0.6$ (see Figure A28). While we can only speculate on the reason for the particular set of patterns across the two precisions, in both cases, the signal composition seems to matter in ways it does not in the uniform treatment.

The takeaway is twofold. First, sequential presentation seems to matter regardless of whether the initial prior is uniform or non-uniform: for sequences common to both treatments, beliefs in the sequential task are more extreme than in the one-shot task. Second, whether the initial signal arrives explicitly or is embedded in the prior has lasting effects on how subsequent information is processed. This suggests that violations of prior sufficiency are driven not only by the possibility of signal aggregation, but also by the initial conditions of the inference problem. Characterizing the relationship between a problem's exogenous parameters and aggregation strategies is a question for future research.

9 Conclusion

This paper experimentally characterizes how individuals update beliefs when information arrives sequentially, focusing on two core implications of Bayesian reasoning in dynamic environments: order independence and prior sufficiency. The experimental literature has largely relied on oneshot inference problems, and, when considering sequences, has focused primarily on order independence. This paper introduces a design that embeds within-person counterfactuals for both properties by combining a sequential updating task with a matched one-shot benchmark.

Across treatments varying prior strength and signal precision, three main findings emerge. First, order dependence is present only for a minority of subjects, who exhibit systematic overreaction to the most recent contradictory signal. Second, prior sufficiency is violated uniformly: posteriors formed endogenously through sequential updating are consistently more extreme than posteriors formed from an informationally equivalent, exogenously given prior. Third, the mechanisms underlying this failure appear linked to aggregate processing of signals. Methodologically, the paper shows that the standard econometric approach of estimating a Grether style AR1 model cannot reconcile observed

patterns, even when estimated flexibly.

The results point to two central directions for further work. The first is about whether the endogeneity of beliefs per se matters. That is, does the source of a belief, not just its numerical value, affect later updating — for example, because of confidence. If yes, then sequential and one-shot tasks correspond to fundamentally different learning problems.

A second avenue concerns whether there exists any static inference problem that, iterated over time, reproduces observed behavior in a sequence. This paper shows that the canonical one-shot task does not serve that role. If no such static analogue exists, sequential reasoning must be modeled as a genuinely dynamic cognitive process.

Figures

Figure 2: Counterfactuals for prior sufficiency

	Sequential problem Worker O				One-shot problem Series of Worker N's		
	Prior	**Signal**	**Posterior**		**Prior**	**Signal**	**Posterior**
$t=1$	p^O	s_1	p_1^O				
$t=2$	p_1^O	s_2	p_2^O	$N1$	$p=p_1^O$	s_2	p^{N1}
$t=3$	p_2^O	s_3	p_3^O	$N2$	$p=p_2^O$	s_3	p^{N2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t=t'$	$p_{t'-1}^O$	$s_{t'}$	$p_{t'}^O$	$N_{t'-1}$	$p=p_{t'-1}^O$	$s_{t'}$	$p^{N_{t'-1}}$

	Sequential problem Worker O				Subsequence problem Single Worker N		
	Prior	**Signal**	**Posterior**		**Prior**	**Signal**	**Posterior**
$t=1$	p^O	s_1	p_1^O				
$t=2$	p_1^O	s_2	p_2^O	N	$p=p_1^O$	s_2	
$t=3$	p_2^O	s_3	p_3^O			s_3	
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	
$t=t'$	$p_{t'-1}^O$	$s_{t'}$	$p_{t'}^O$			$s_{t'}$	$p_{t'}^N$

Order independence in the aggregate

Figure 3: CDF: Posterior after sequences with composition {0,1}
$p = 0.5$, $q = 0.6$

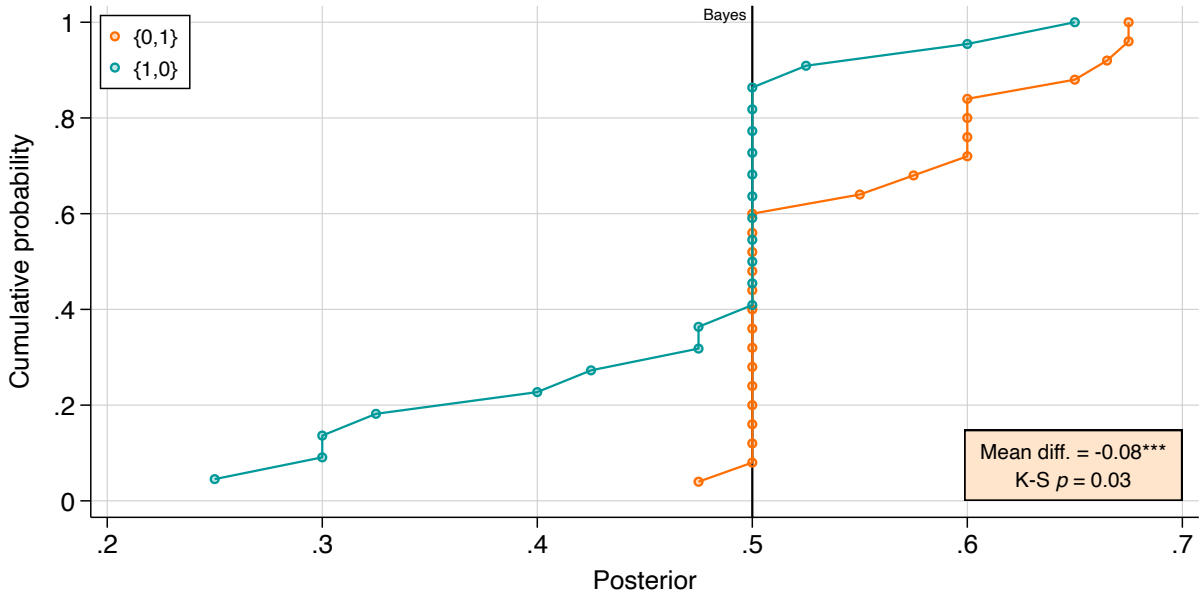
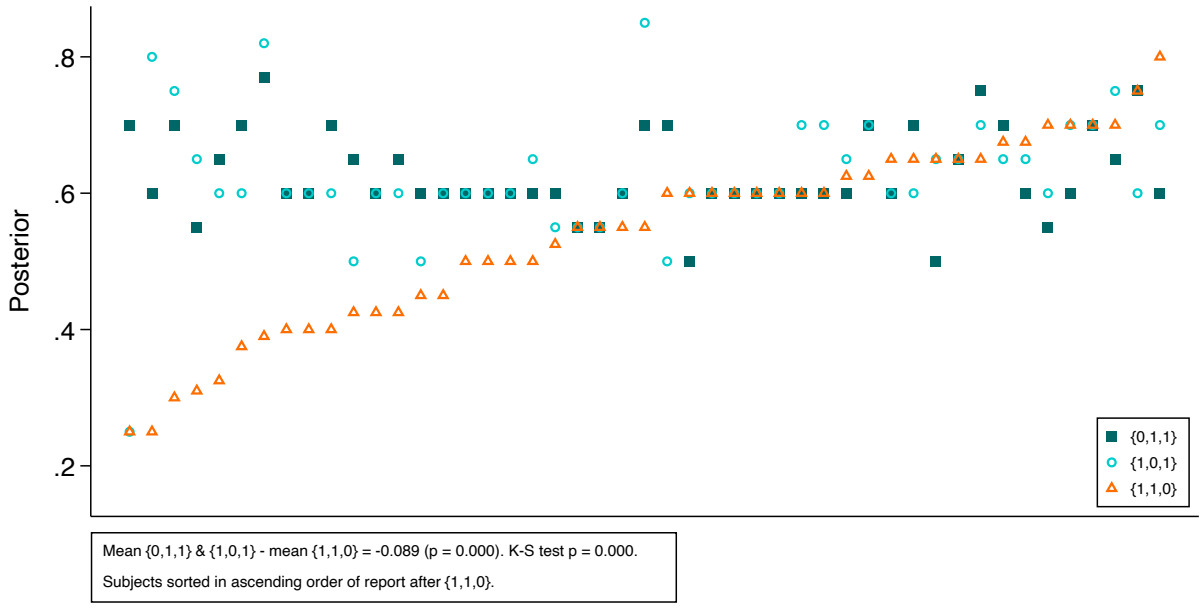


Figure 4: Posterior after sequences with composition {0,1,1}, at subject level $p = 0.5$, $q = 0.6$



Prior sufficiency in the aggregate

Figure 5: Average posterior after $\{1,0,1,1\}$ & $\{1,1,0,1\}$, at subject level
$p = 0.5$, $q = 0.6$

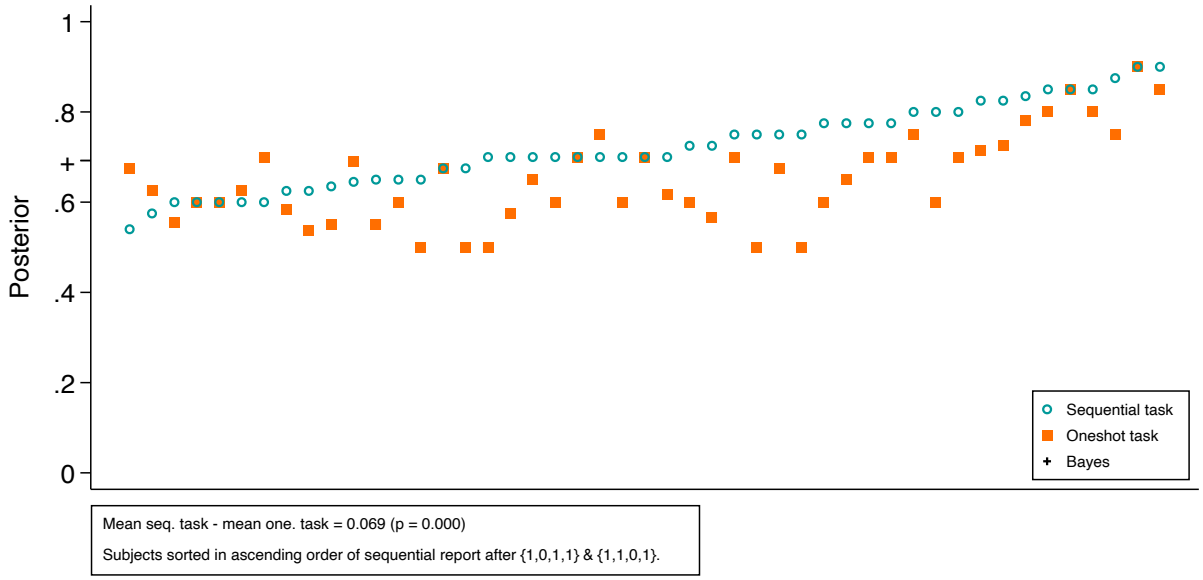
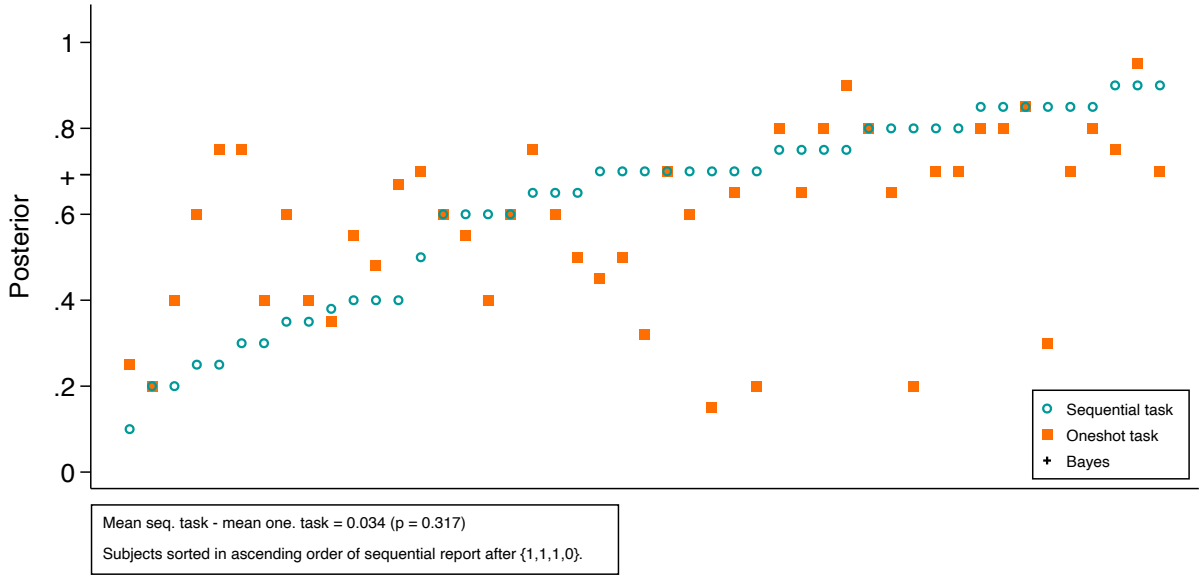
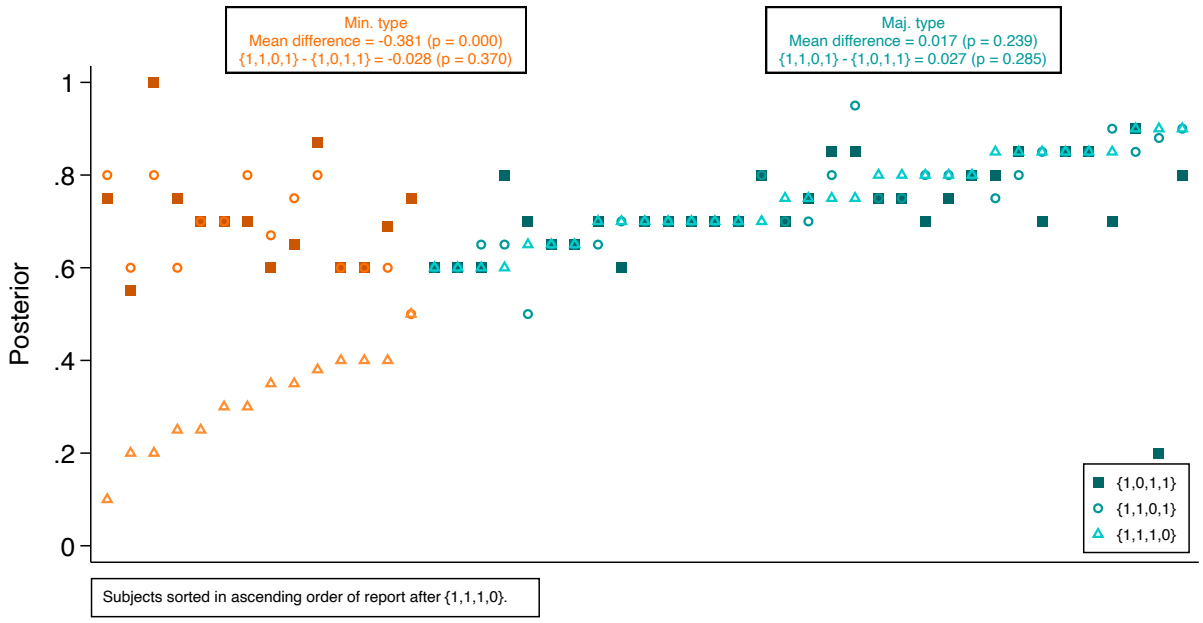


Figure 6: Posterior after $\{1,1,1,0\}$, at subject level
$p = 0.5$, $q = 0.6$



Heterogeneity in violations: order independence

Figure 7: Posterior after sequences with composition {0,1,1,1}, at subject level $p = 0.5$, $q = 0.6$



Heterogeneity in violations: prior sufficiency

Figure 8: Avg. report by task and time
Sequence: {1,0,1,1}
$p = 0.5, q = 0.6$; Maj. type

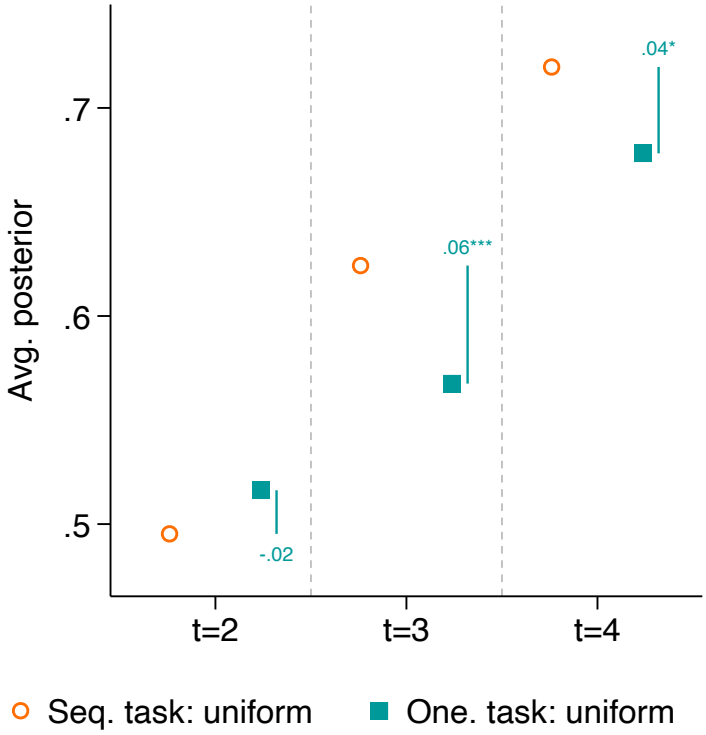


Figure 9: Avg. report by task and time
Sequence: {1,1,0,1}
$p = 0.5, q = 0.6$; Maj. type

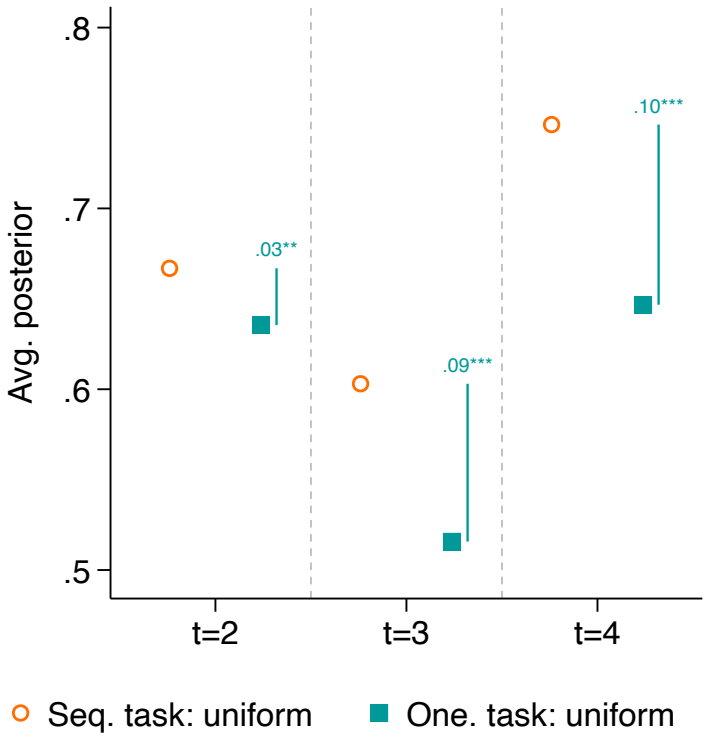


Figure 10: Avg. report by task and time
Sequence: {1,1,1,0}
$p = 0.5, q = 0.6$; Maj. type

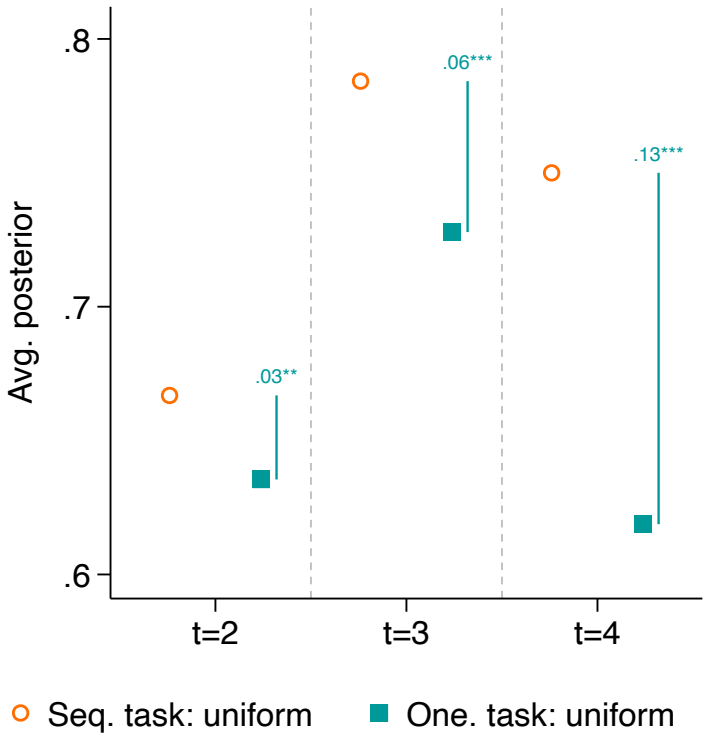


Figure 11: CDF: Posterior in sequential task vs. oneshot task
Sequence {1,0,1,1} & {1,1,0,1}
$p = 0.5$, $q = 0.6$; Maj. type

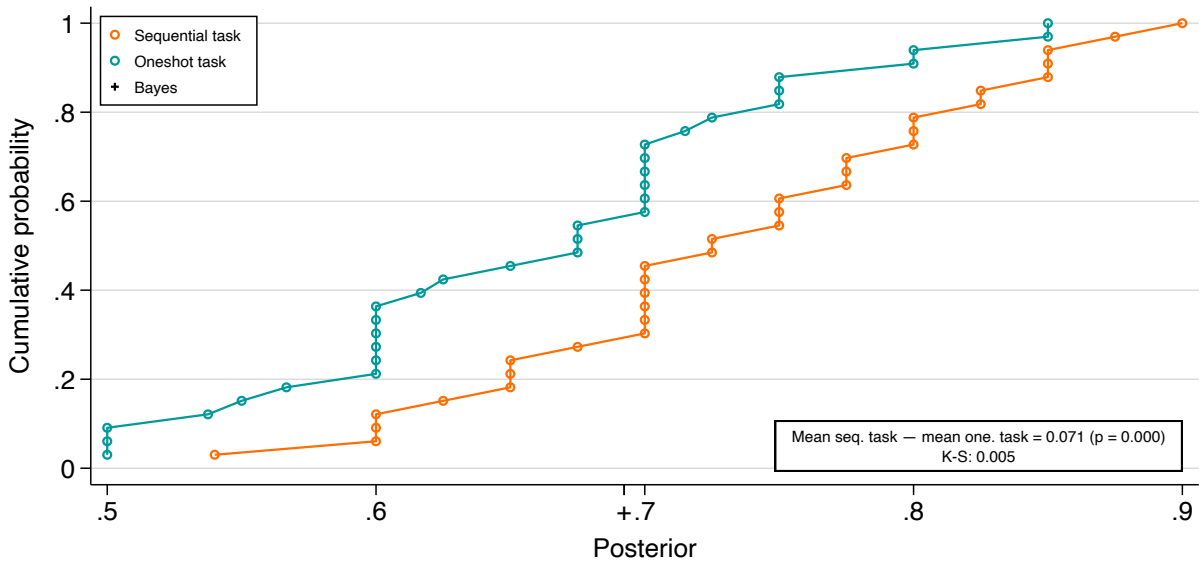


Figure 12: CDF: Posterior in sequential task vs. oneshot task
Sequence {1,1,1,0}
$p = 0.5$, $q = 0.6$; Maj. type

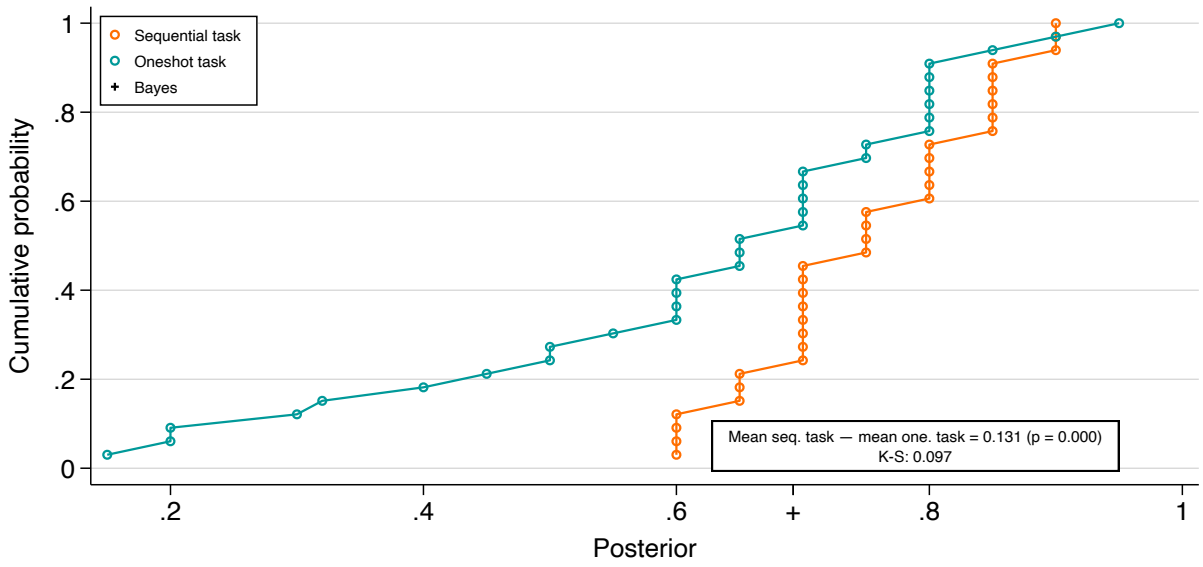


Figure 13: CDF: Posterior in sequential task vs. oneshot task
Sequence {1,1,0,1}
$p = 0.5$, $q = 0.6$; Min. type

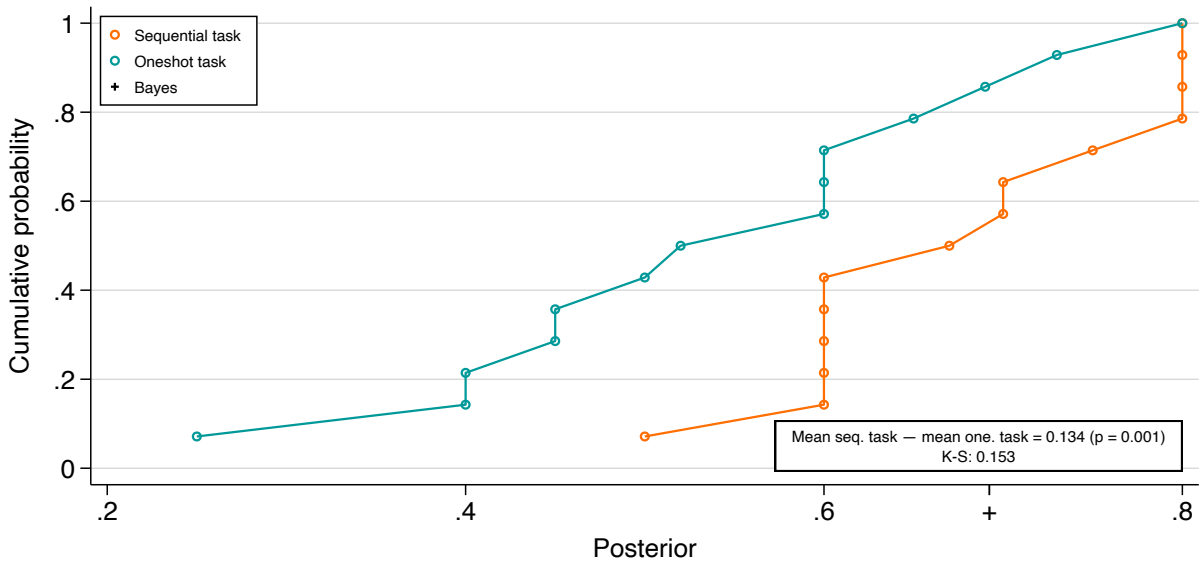
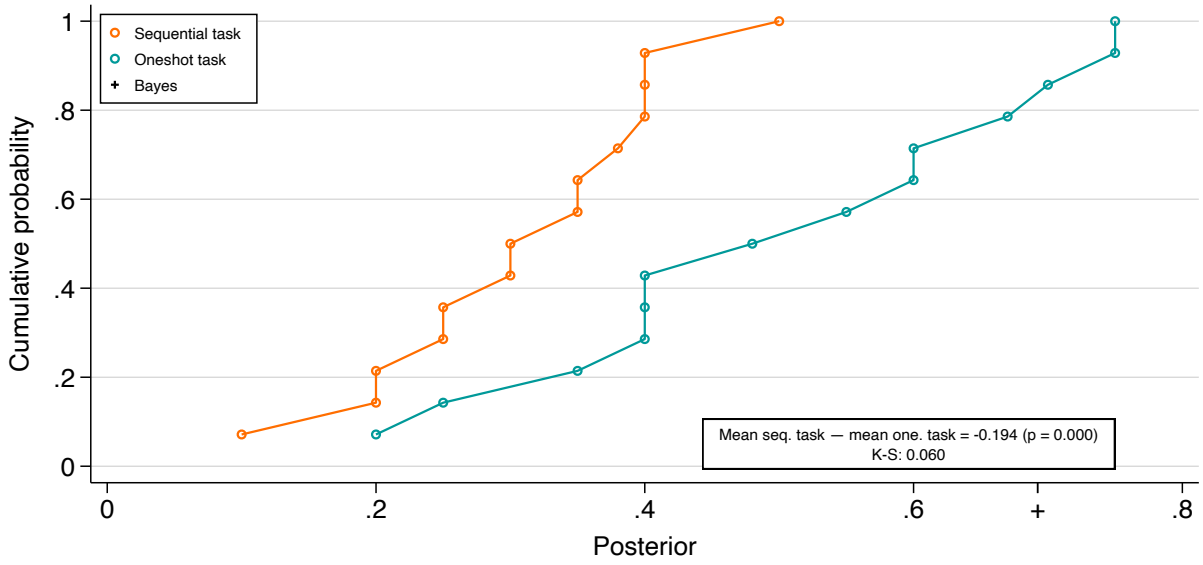


Figure 14: CDF: Posterior in sequential task vs. oneshot task
Sequence {1,1,1,0}
$p = 0.5$, $q = 0.6$; Min. type



Prior sufficiency with respect to a sequence

Figure 15: Avg. report by task and time
Sequence: {1,0,1,1}
$q = 0.6$

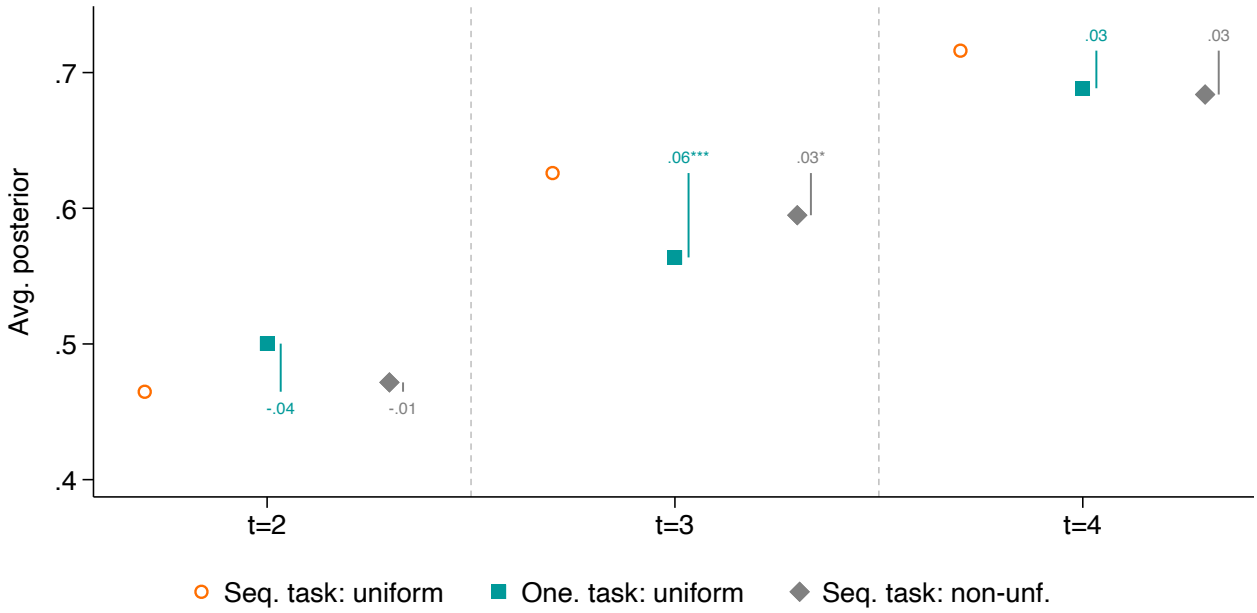


Figure 16: Avg. report by task and time
Sequence: {1,1,0,1}
$q = 0.6$

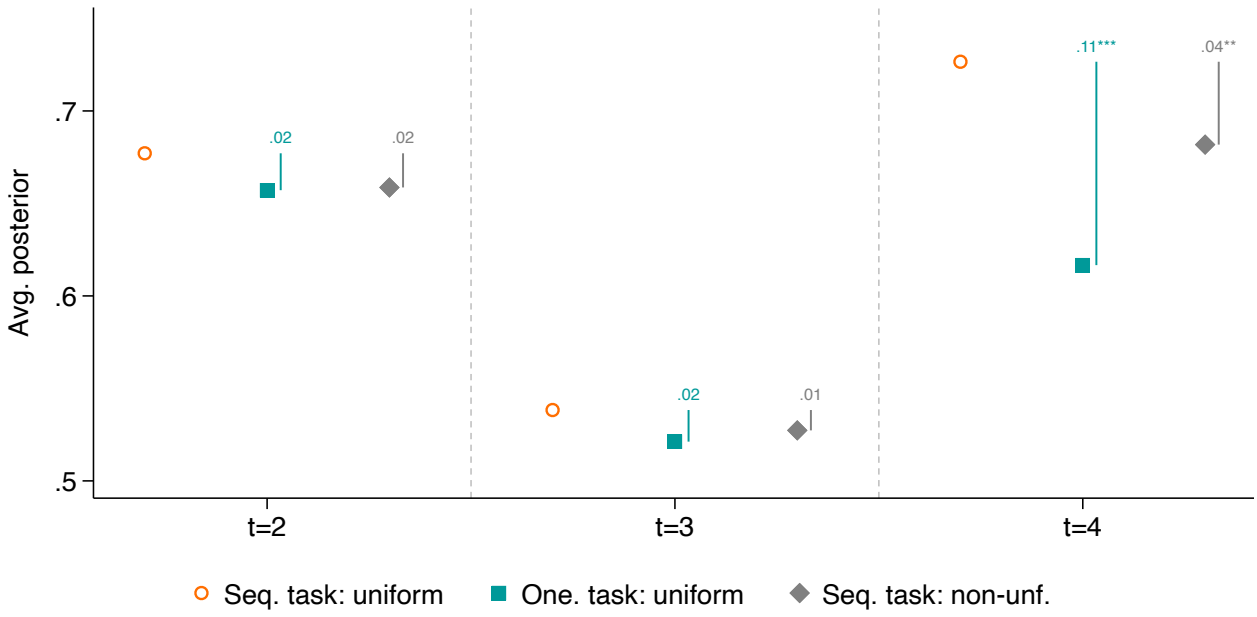


Figure 17: Avg. report by task and time
Sequence: {1,1,1,0}
$q = 0.6$

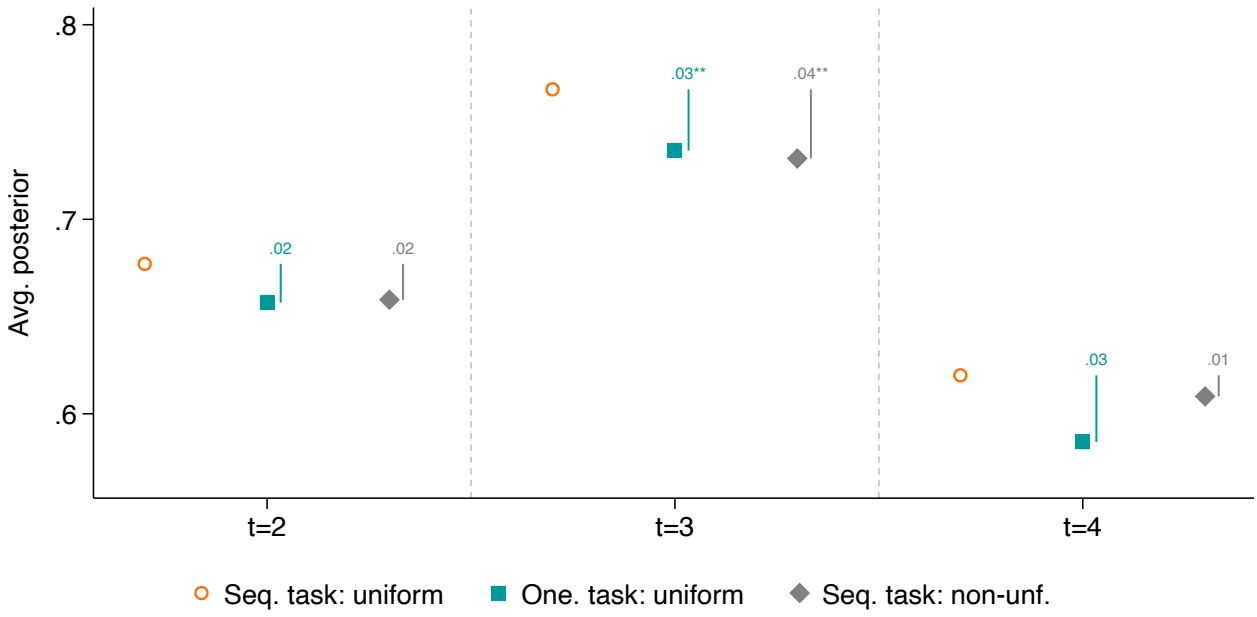


Figure 18: CDF: Posterior in sequential task, by prior
Sequence {1,0,1,1} & {1,1,0,1}
$q = 0.8$

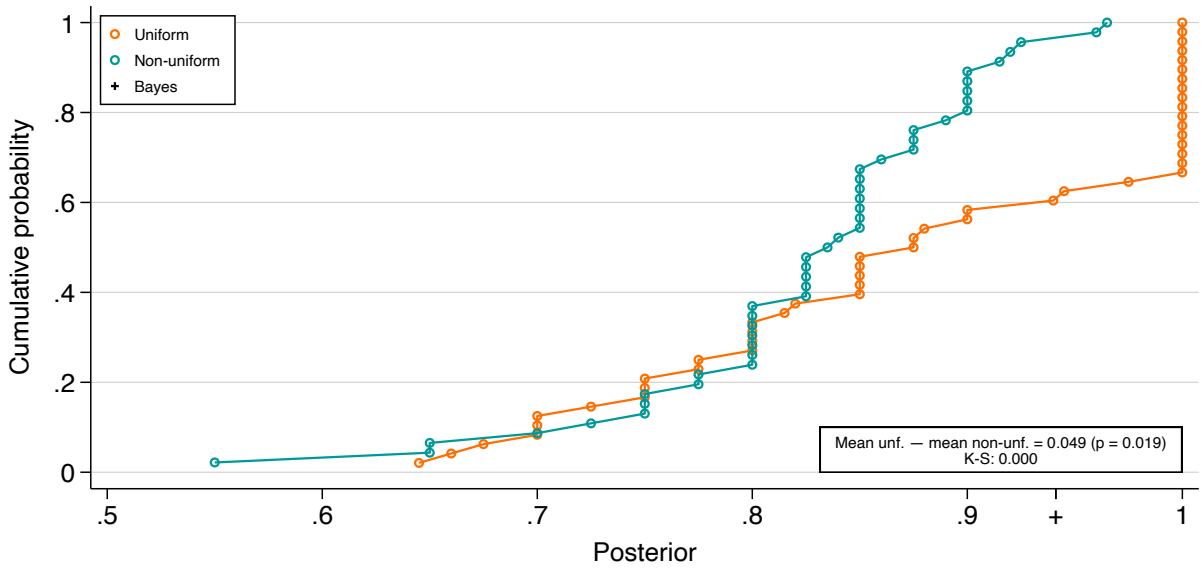


Figure 19: CDF: Posterior in sequential task, by prior
Sequence {1,1,1,0}
$q = 0.8$

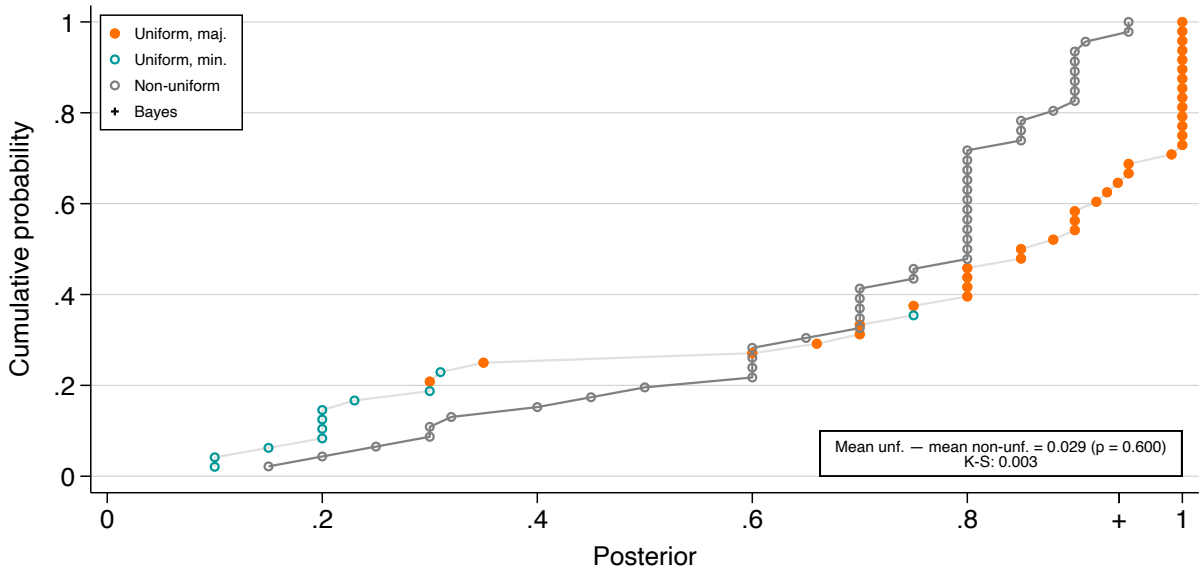
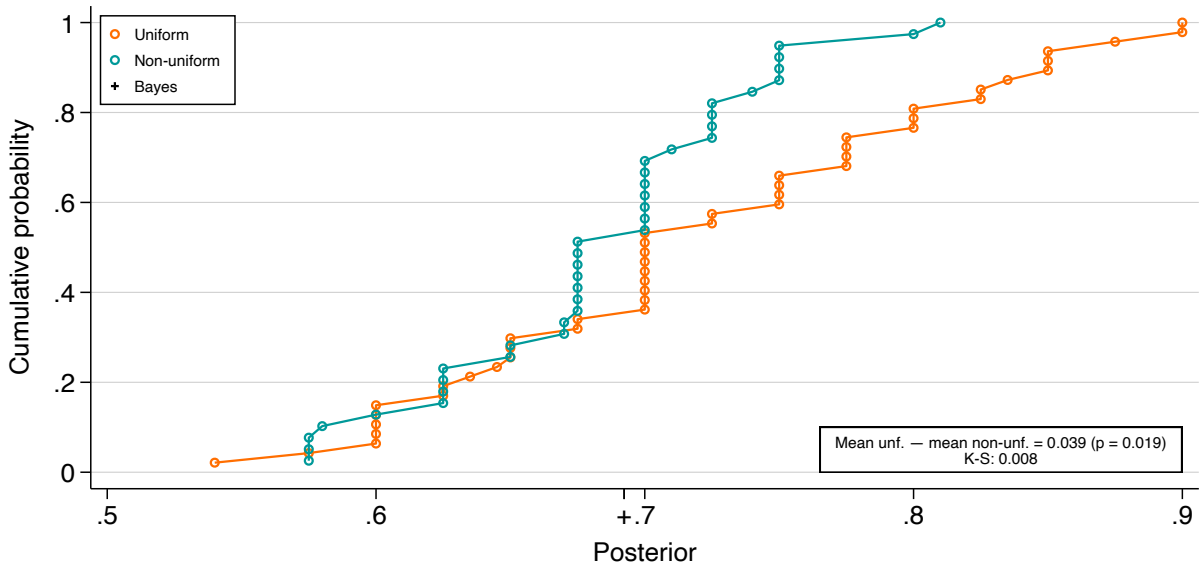
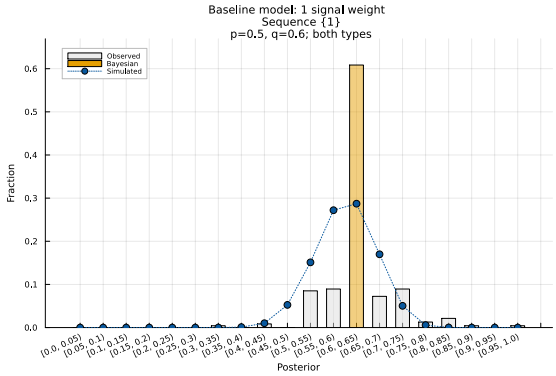


Figure 20: CDF: Posterior in sequential task, by prior
Sequence {1,0,1,1} & {1,1,0,1}
$q = 0.6$

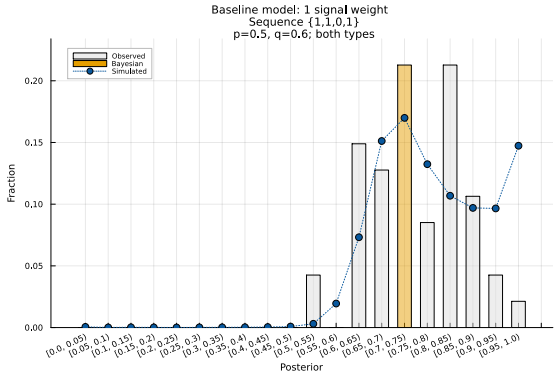


Estimation of updating rules

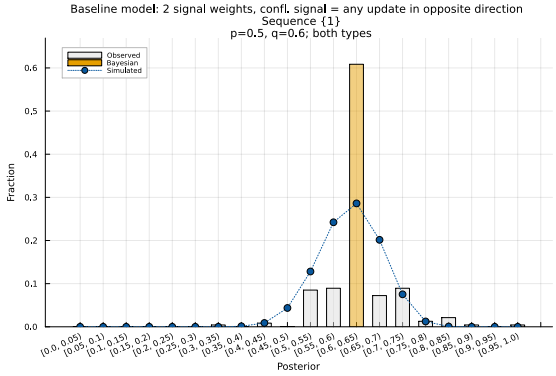
Figure 21: Observed data versus data simulated under Grether model
Treatment: $p = 0.5, q = 0.6$



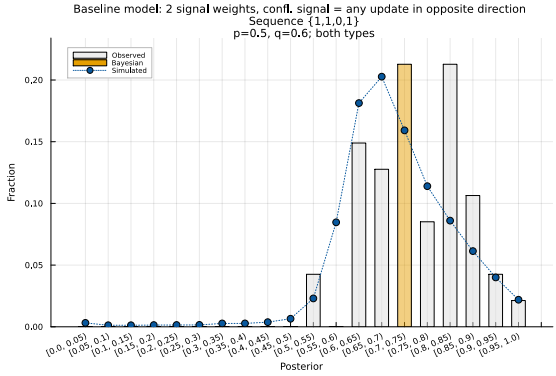
(a)



(b)

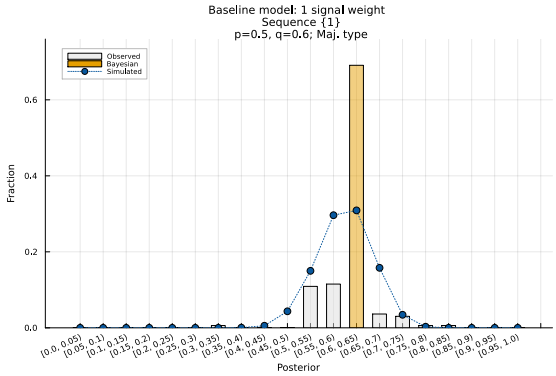


(c)

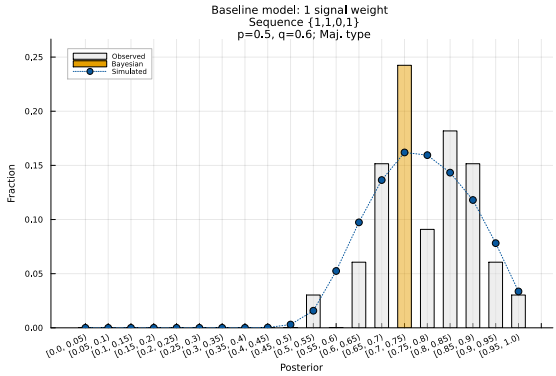


(d)

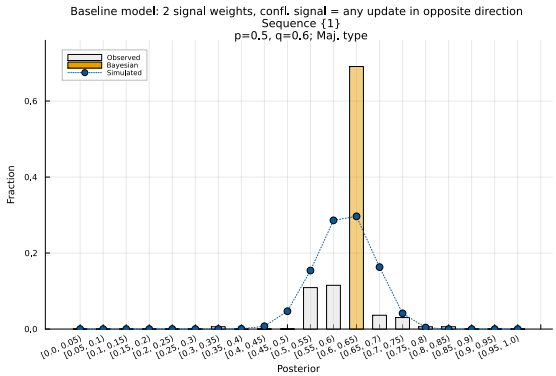
Figure 22: Observed data versus data simulated under Grether model
Treatment: $p = 0.5, q = 0.6$
Maj. type



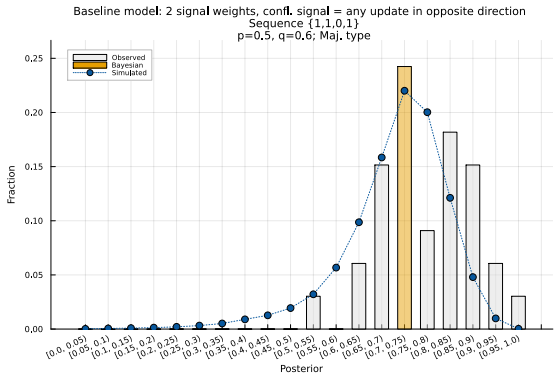
(a)



(b)



(c)



(d)

Sources of prior sufficiency violations: aggregating signals

Figure 23: CDF: Posterior in sequential task
Sequences that reduce to two 1 signals
$p = 0.5$, $q = 0.6$; Maj. type

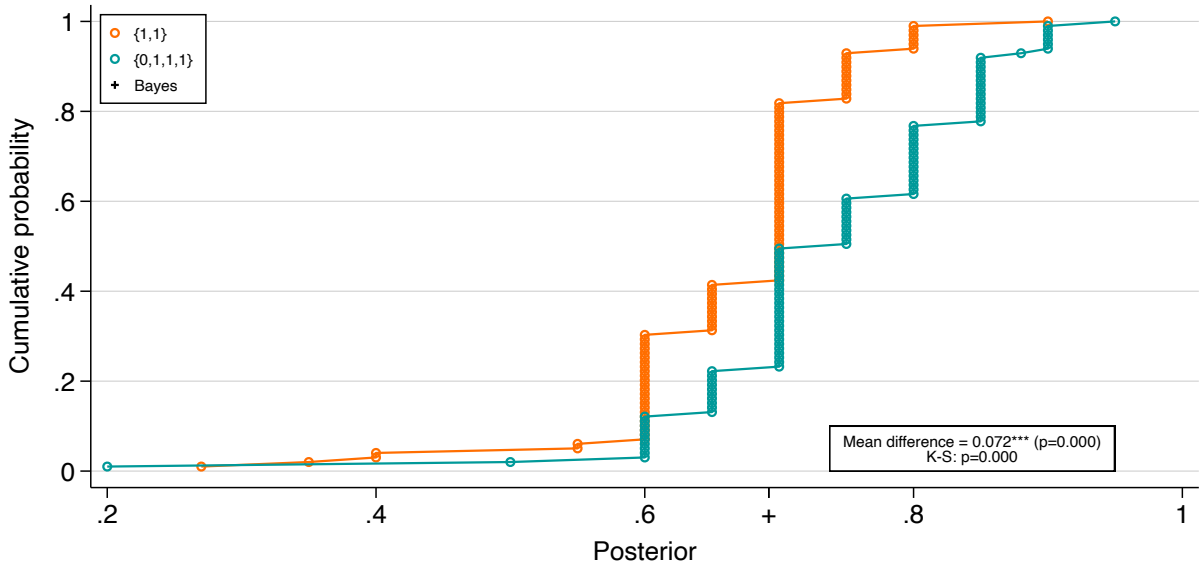
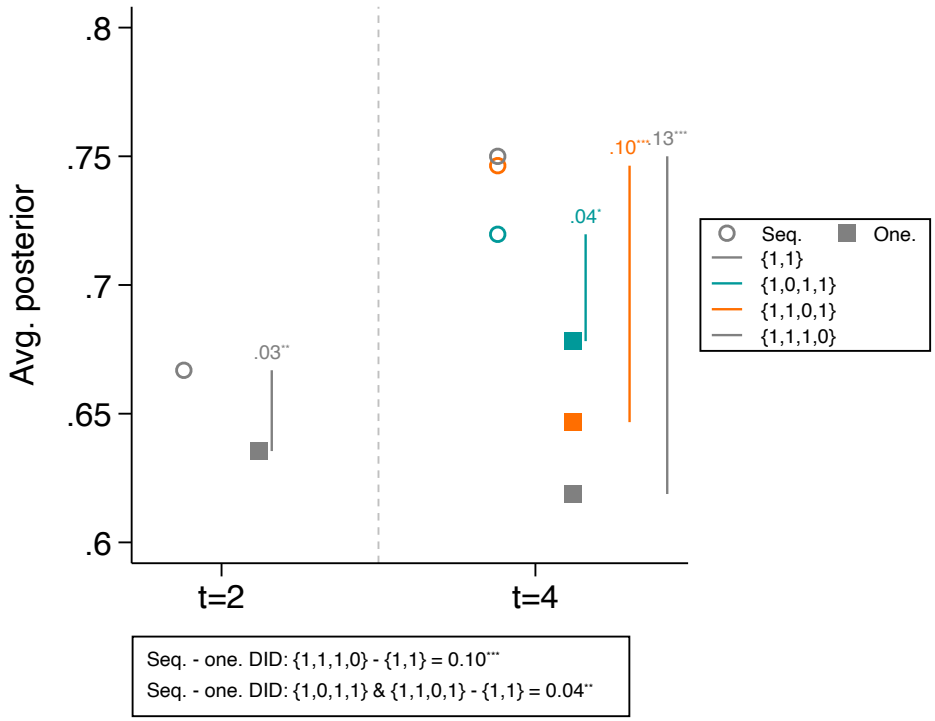


Figure 24: Avg. posterior by time
Sequences with two 1 signals
$p = 0.5$, $q = 0.6$; Maj. type



Sources of prior sufficiency violations: updating in the non-uniform treatments

Figure 25: CDF: Posterior in sequential vs. one-shot task
Sequences {1,0,1,1} & {1,1,0,1}
$p = 0.6$, $q = 0.6$

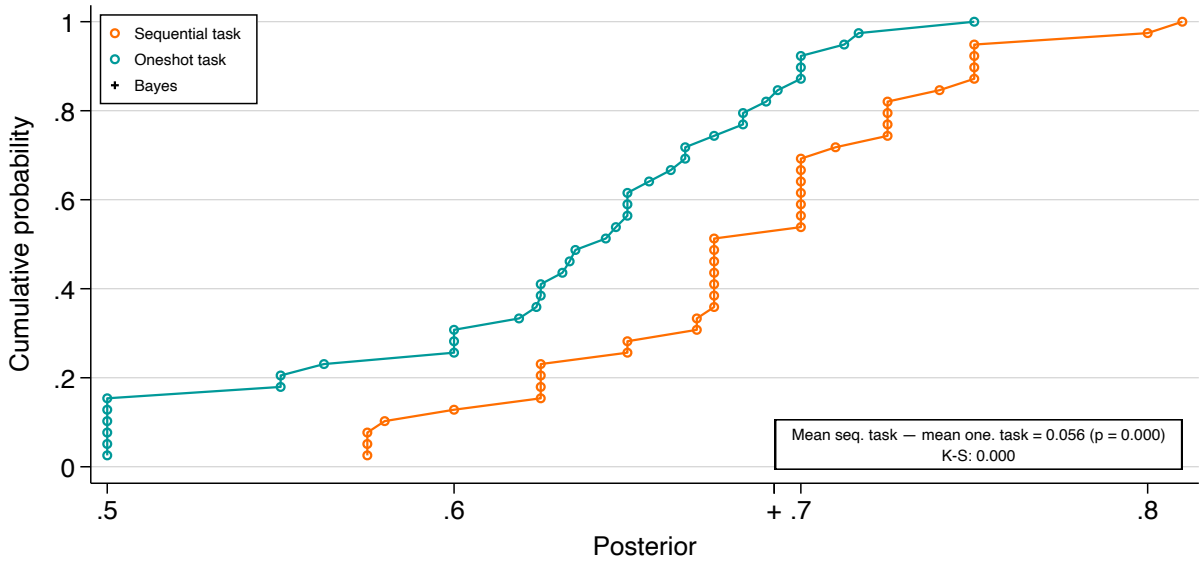


Figure 26: CDF: Posterior in sequential vs. one-shot task
Sequences of length 3 with one vs. two 1 signals
$p = 0.6$, $q = 0.6$

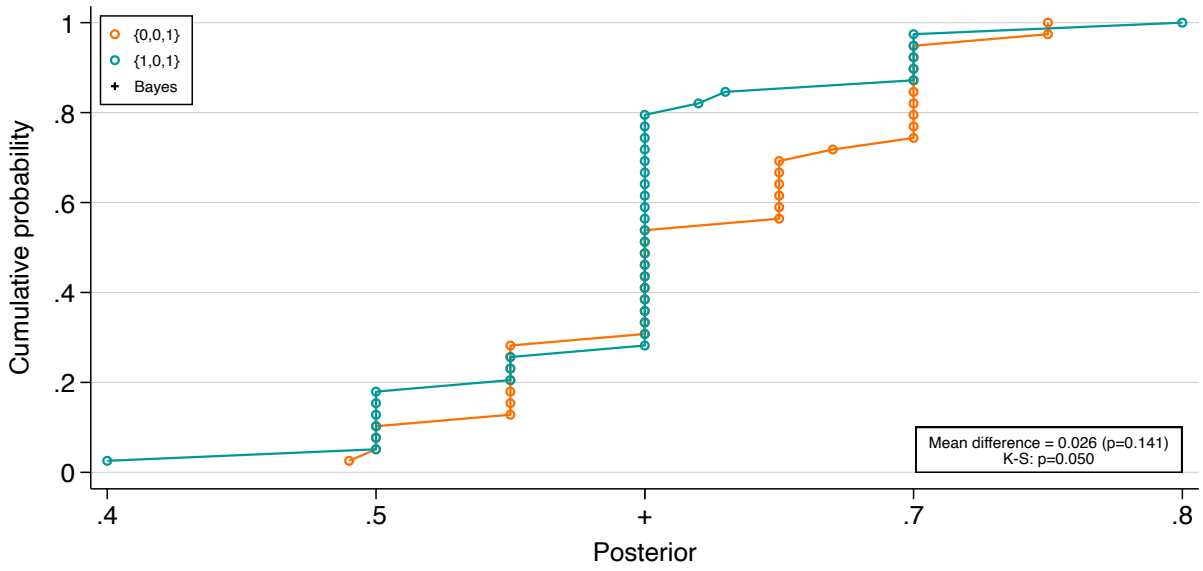
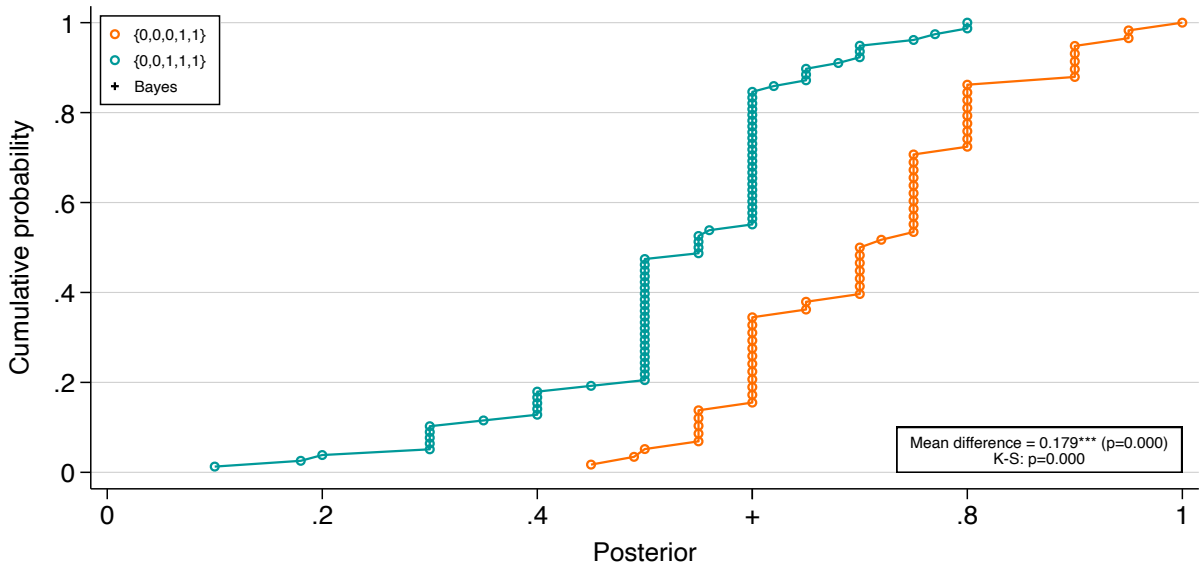


Figure 27: CDF: Posterior in sequential vs. one-shot task
Sequences of length 5 with two vs. three 1 signals
$p = 0.6$, $q = 0.6$



Tables

Table 1: Treatments

	Prior p	**Precision q**	**Total # Participants**	**# Participants in Main Sample**
1	.5	.6	60	47
2	.5	.8	60	48
3	.6	.6	60	39
4	.8	.8	60	46

Table 2: Sequences

(A) $p = .5$, $q = .6$

$t = 1$	$t = 2$	$t = 3$	$t = 4$
1	0	1	1
1	1	0	1
1	1	1	0
1	0	0	1
1	1	0	0

(B) $p = .6$, $q = .6$

$t = 1$	$t = 2$	$t = 3$	$t = 4$
0	1	1	1
1	0	1	1
1	1	0	0
0	0	1	1
1	0	0	0

(C) $p = .5$, $q = .8$

$t = 1$	$t = 2$	$t = 3$	$t = 4$
1	0	1	1
1	1	0	1
1	1	1	0
1	0	0	1
1	1	1	1

(D) $p = .8$, $q = .8$

$t = 1$	$t = 2$	$t = 3$	$t = 4$
0	1	1	1
1	0	1	1
1	1	0	0
0	0	1	1
1	1	1	1

Table 3: Match on modal bins for baseline models
Treatment: $p = 0.5, q = 0.6$

Sequence	Both types baseline 1 wgt	Both types baseline 2 wgt	Maj. type baseline 1 wgt	Maj. type baseline 2 wgt
{1}	0.322	0.321	0.383	0.392
{1,0}	0.314	0.299	-0.194	0.090
{1,1}	0.134	0.138	-0.117	0.202
{0,1,1}	0.311	0.285	0.223	0.287
{1,0,1}	0.100	0.131	0.143	0.077
{1,1,0}	0.199	0.157	0.112	0.087
{1,1,1}	0.204	0.212	0.261	0.258
{1,0,1,1}	0.129	0.045	0.052	0.122
{1,1,0,1}	0.042	0.054	0.081	0.025
{1,1,1,0}	0.117	0.085	0.098	0.053
{0,1,1,0}	0.475	0.379	0.229	0.489
{1,1,0,0}	0.528	0.487	0.569	0.444

Notes: Each column corresponds to a sample and model. Each row shows the difference between the fraction of observed data in the observed modal bin and the fraction of simulated data in the observed modal bin.

Table 4: Match on non-modal bins for baseline models
Treatment: $p = 0.5, q = 0.6$

Sequence	Both types baseline 1 wgt	Both types baseline 2 wgt	Maj. type baseline 1 wgt	Maj. type baseline 2 wgt
{1}	0.242	0.208	0.207	0.210
{1,0}	0.264	0.221	0.097	0.233
{1,1}	0.226	0.234	0.285	0.244
{0,1,1}	0.259	0.293	0.325	0.290
{1,0,1}	0.165	0.232	0.317	0.244
{1,1,0}	0.238	0.242	0.197	0.279
{1,1,1}	0.215	0.218	0.203	0.192
{1,0,1,1}	0.225	0.224	0.236	0.289
{1,1,0,1}	0.253	0.223	0.141	0.258
{1,1,1,0}	0.301	0.225	0.139	0.244
{0,1,1,0}	0.365	0.273	0.314	0.478
{1,1,0,0}	0.353	0.311	0.315	0.250

Notes: Each column corresponds to a sample and model. Each row shows the absolute difference between the fraction of observed data and simulated data in each non-modal bin, summed over bins.

Table 5: Match on mode and non-modal bins for baseline model with 1 vs 2 signal weights
Treatment: $p = 0.5, q = 0.6$

	Both types		Maj. type	
Sequence	2 −1 mode	2 −1 nonmode	2 −1 mode	2 −1 nonmode
{1}	-0.000	-0.067	0.009	0.006
{1,0}	-0.016	-0.087	0.284	0.272
{1,1}	0.004	0.015	0.319	-0.083
{0,1,1}	-0.025	0.069	0.064	-0.070
{1,0,1}	0.030	0.135	-0.066	-0.147
{1,1,0}	-0.042	0.007	-0.025	0.164
{1,1,1}	0.008	0.007	-0.003	-0.023
{1,0,1,1}	-0.084	-0.002	0.071	0.107
{1,1,0,1}	0.012	-0.061	-0.056	0.234
{1,1,1,0}	-0.032	-0.151	-0.045	0.210
{0,1,1,0}	-0.096	-0.183	0.259	0.327
{1,1,0,0}	-0.041	-0.084	-0.125	-0.129

Notes: Each column shows the difference between the 2-weight and 1-weight models. Gray cells are cases where the 2-weight model performs better.

Table 6: Sequences in non-uniform $q = 0.6$ treatment that are equivalent after inversion

Comp.	**Sequences**
2 0s, 1 1	{0,0,1}
1 0, 2 1s	{1,0,1}, {1,1,0}
3 0s, 2 1s	{1,0,0,1,0}, {1,1,0,0,0}
3 1s, 2 0s	{1,0,0,1,1}, {1,0,1,1,0}, {1,1,0,1,0}, {1,1,1,0,0}

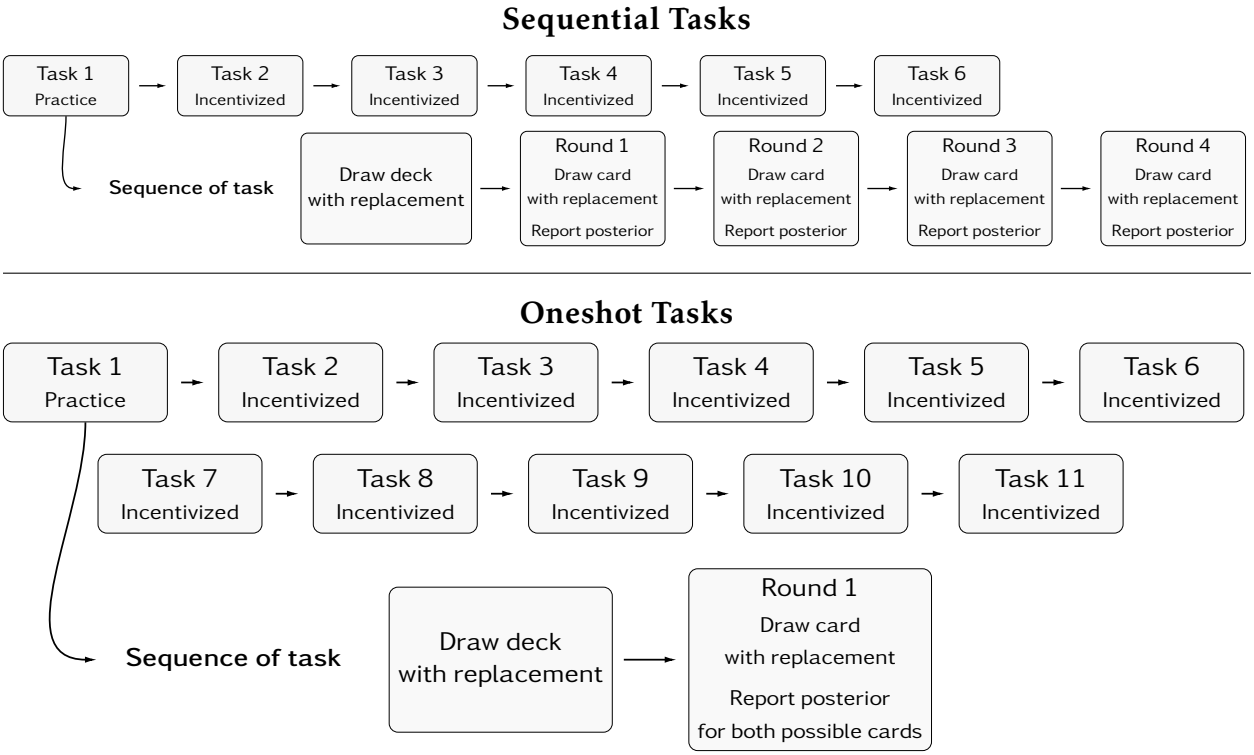
Table 7: Sequences in non-uniform $q = 0.8$ treatment that are equivalent after inversion

Comp.	**Sequences**
2 0s, 1 1	{0,0,1}
1 0, 2 1s	{1,0,1}, {1,1,0}
3 0s, 1 1	{1,0,0,0}
1 0, 3 1s	{1,0,1,1}, {1,1,0,1}, {1,1,1,0}
3 0s, 2 1s	{1,0,0,1,0}
3 1s, 2 0s	{1,0,1,1,0}, {1,1,1,0,0}
4 0s, 1 1	{1,0,0,0,0}
1 0, 4 1s	{1,0,1,1,1}, {1,1,0,1,1}, {1,1,1,0,1}, {1,1,1,1,0}

Appendix

A Figures

Figure A1: Timeline of Study



Order independence in the aggregate

Figure A2: Posterior after sequences with composition {0,1,1,1}, at subject level
$p = 0.5$, $q = 0.6$

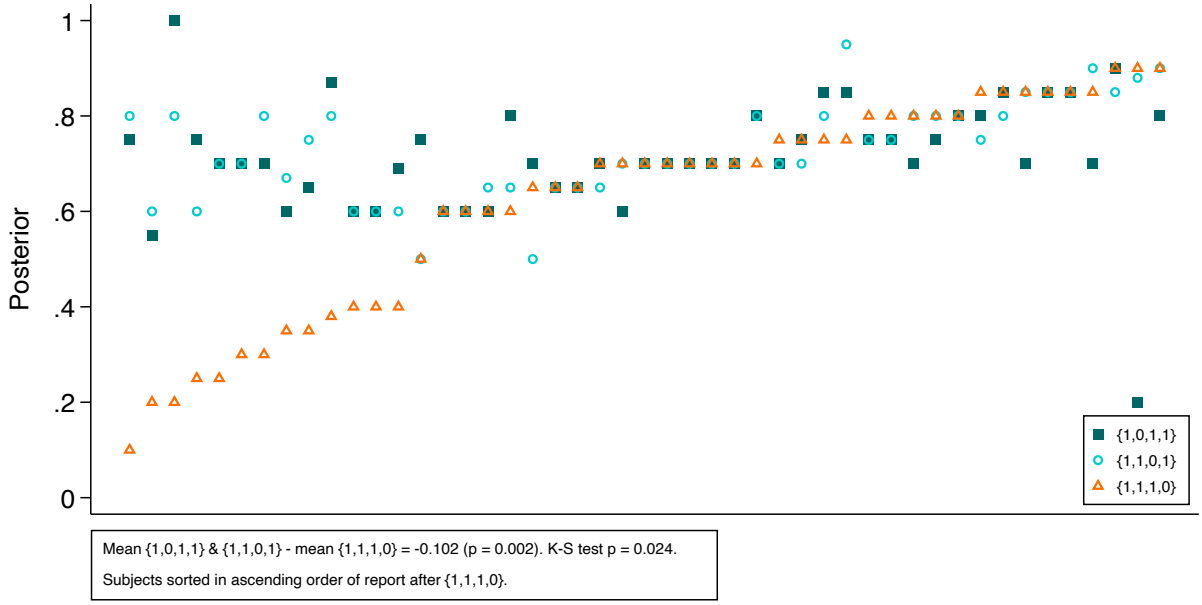
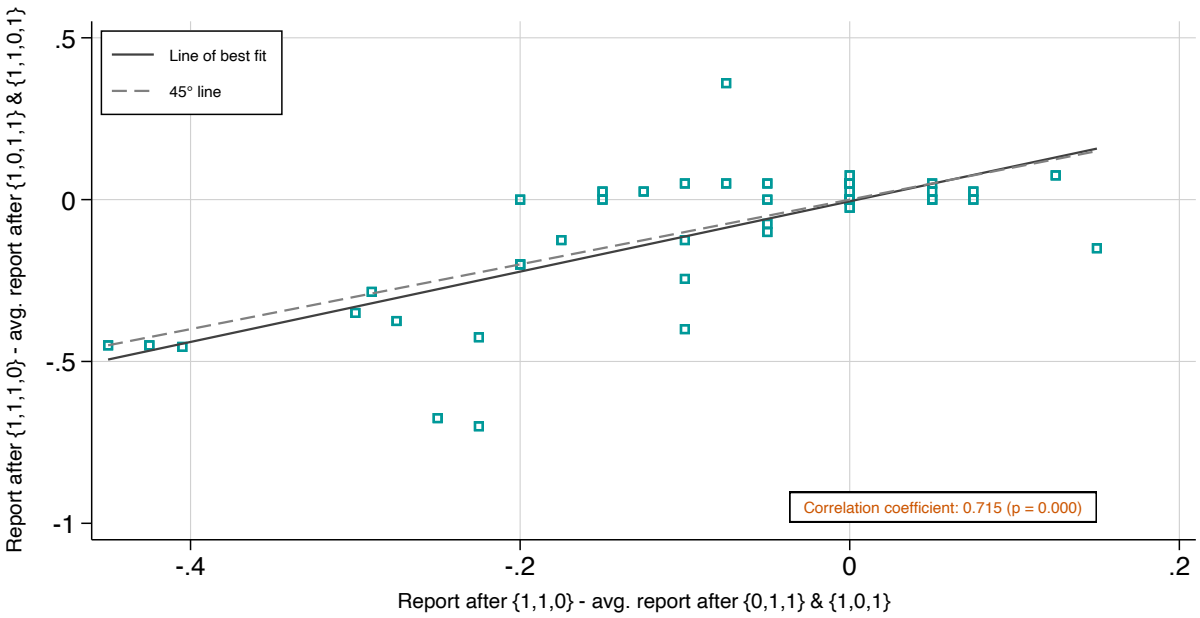
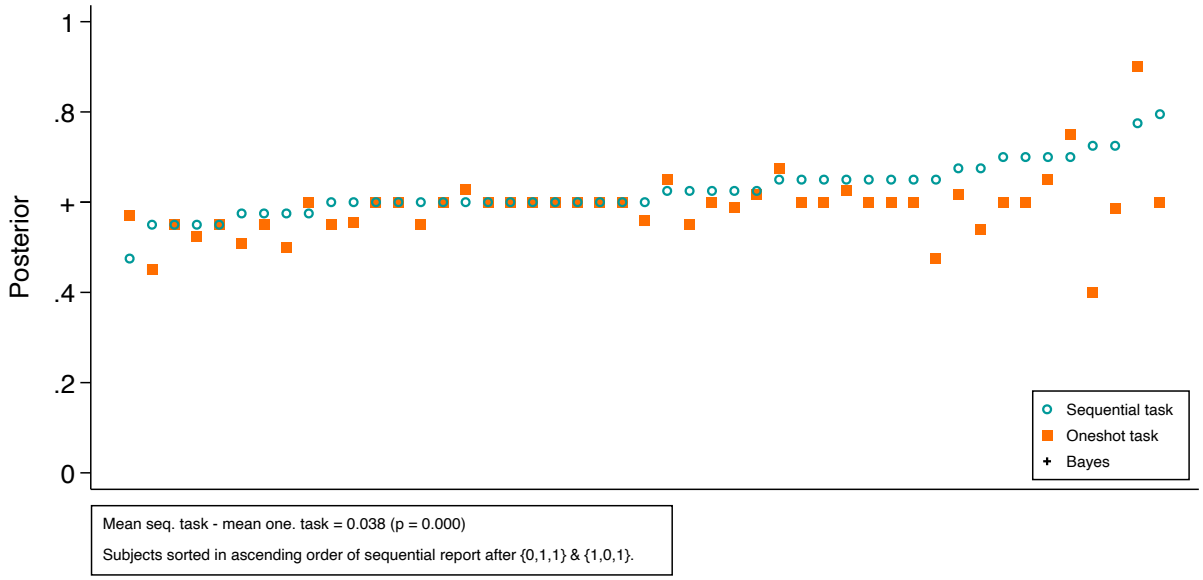


Figure A3: Within subject order effect at $t = 4$ vs. $t = 3$
$p = 0.5$, $q = 0.6$



Prior sufficiency in the aggregate

Figure A4: Average posterior after $\{0,1,1\}$ & $\{1,0,1\}$, at subject level
$p = 0.5$, $q = 0.6$



Heterogeneity in violations: order independence

Figure A5: Posterior after sequences with composition {0,1,1}, at subject level $p = 0.5$, $q = 0.6$

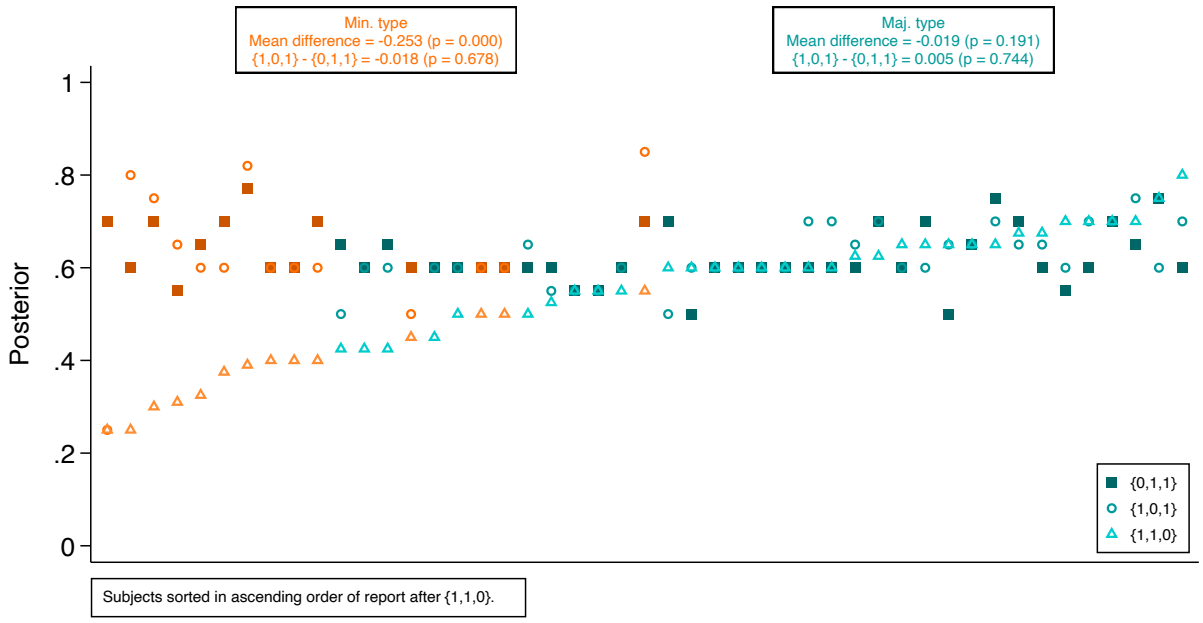
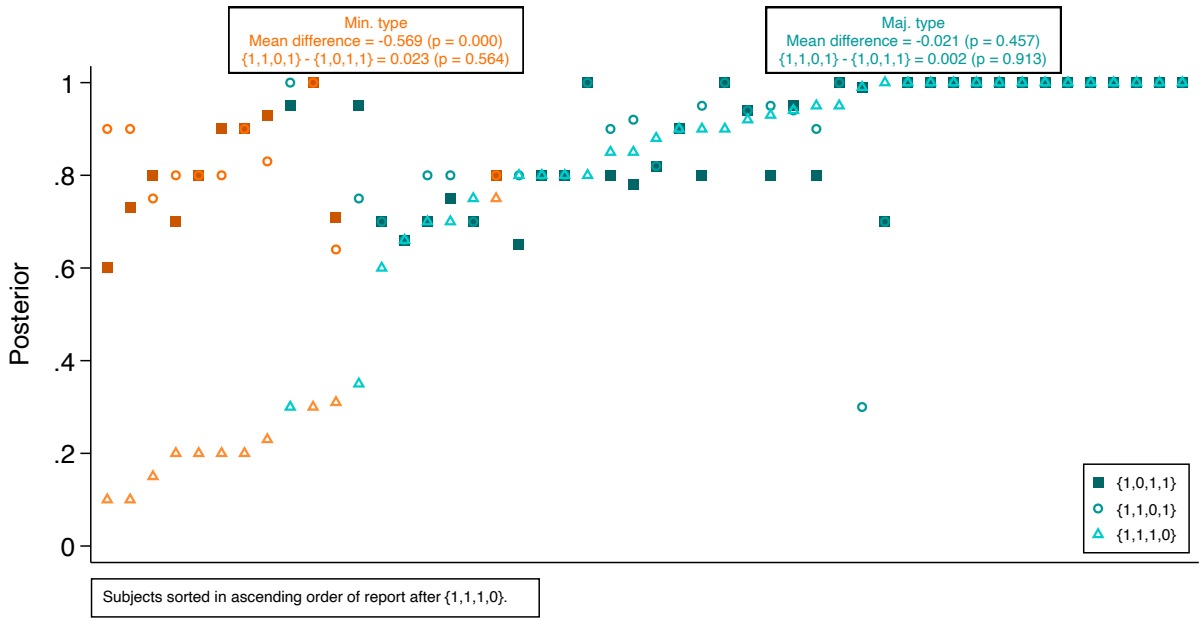


Figure A6: Posterior after sequences with composition {0,1,1,1}, at subject level
$p = 0.5$, $q = 0.8$



Heterogeneity in violations: prior sufficiency

Figure A7: CDF: Posterior in sequential task vs. oneshot task
Sequence {0,1,1} & {1,0,1}
$p = 0.5$, $q = 0.6$; Maj. type

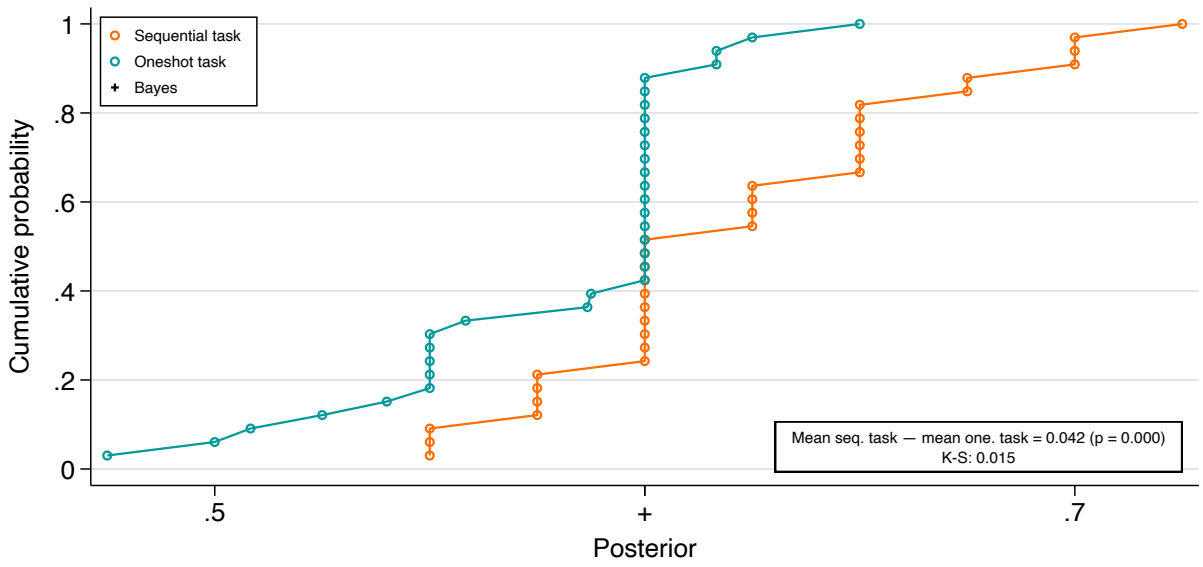


Figure A8: CDF: Posterior in sequential task vs. oneshot task
Sequence {1,1,0}
$p = 0.5$, $q = 0.6$; Maj. type

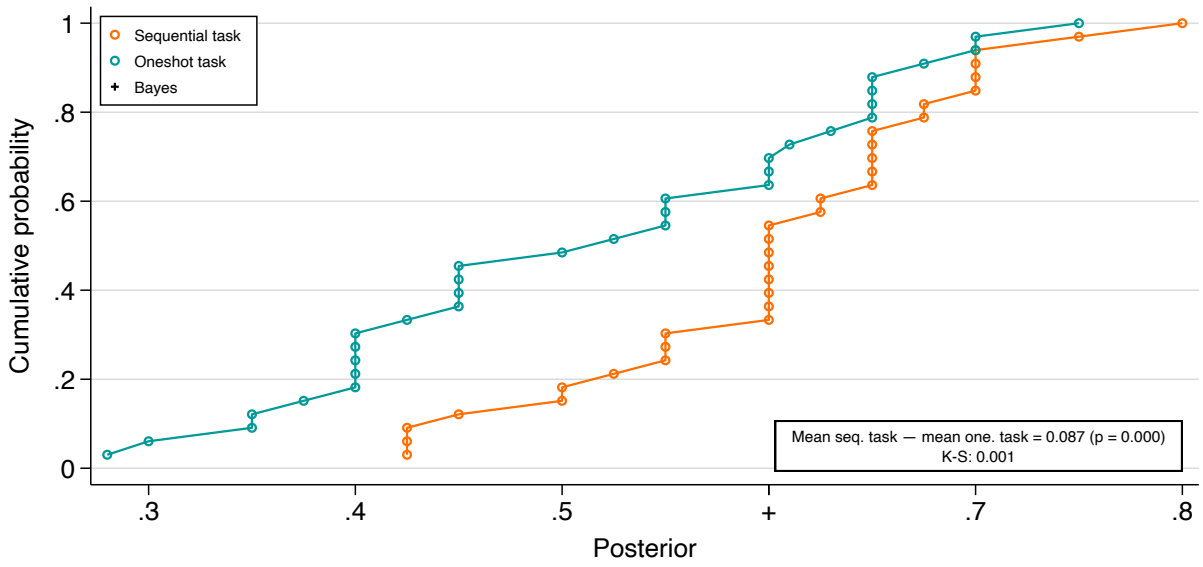


Figure A9: CDF: Posterior in sequential task - posterior in oneshot task, within subject $p = 0.5$, $q = 0.6$; Maj. type

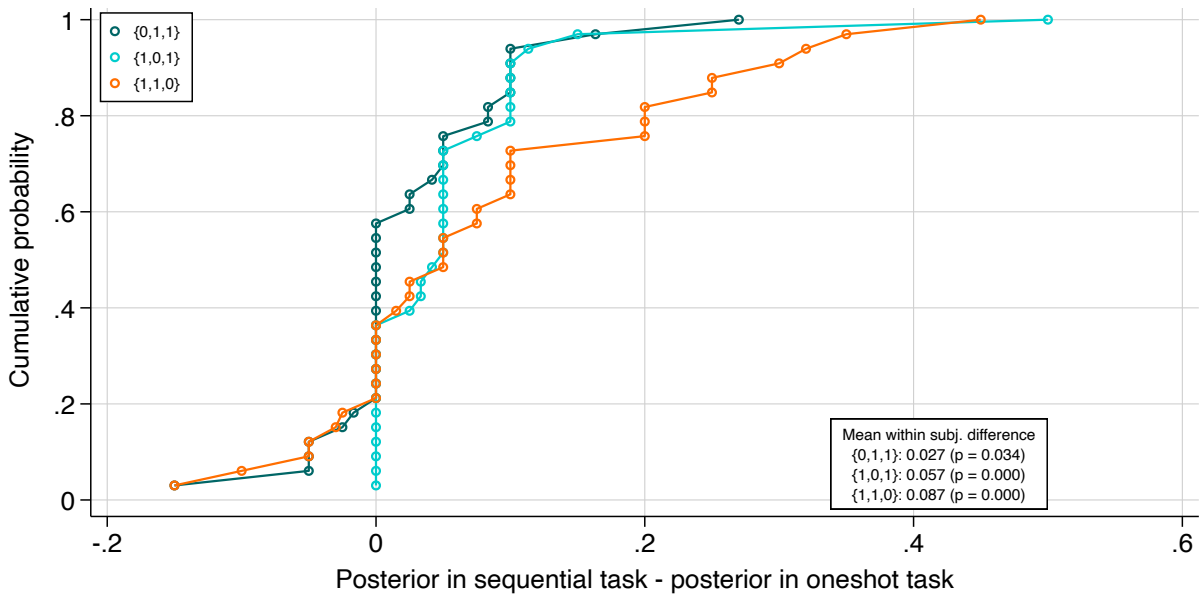
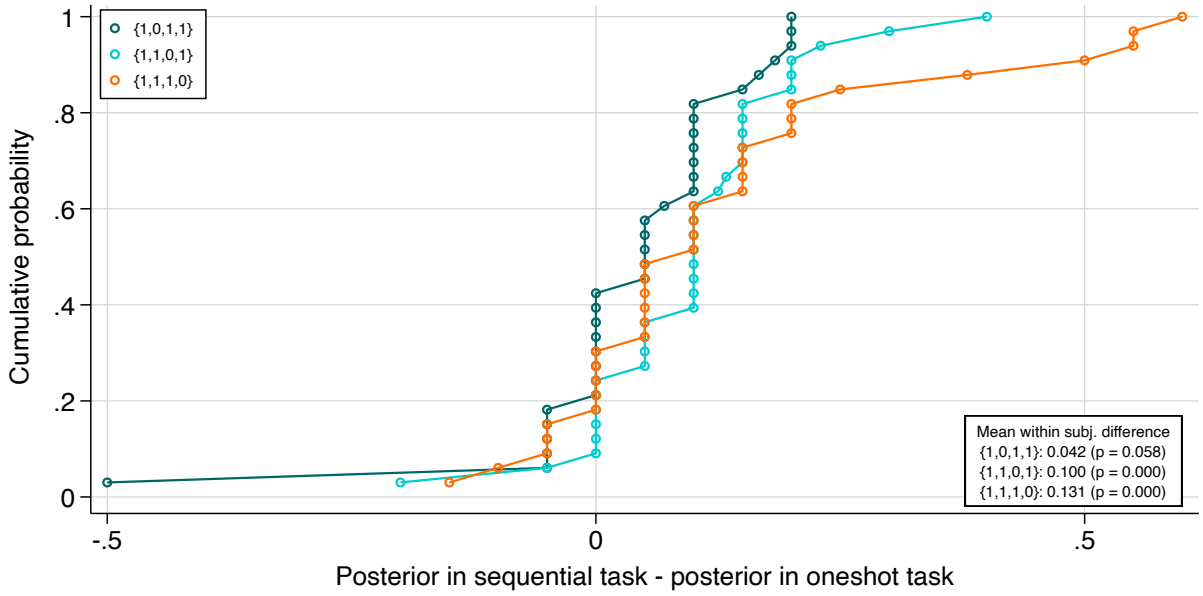


Figure A10: CDF: Posterior in sequential task - posterior in oneshot task, within subject
$p = 0.5$, $q = 0.6$; Maj. type



Prior sufficiency with respect to a sequence

Figure A11: Avg. report by task and time
Sequence: {0,1,1,0}
$q = 0.6$

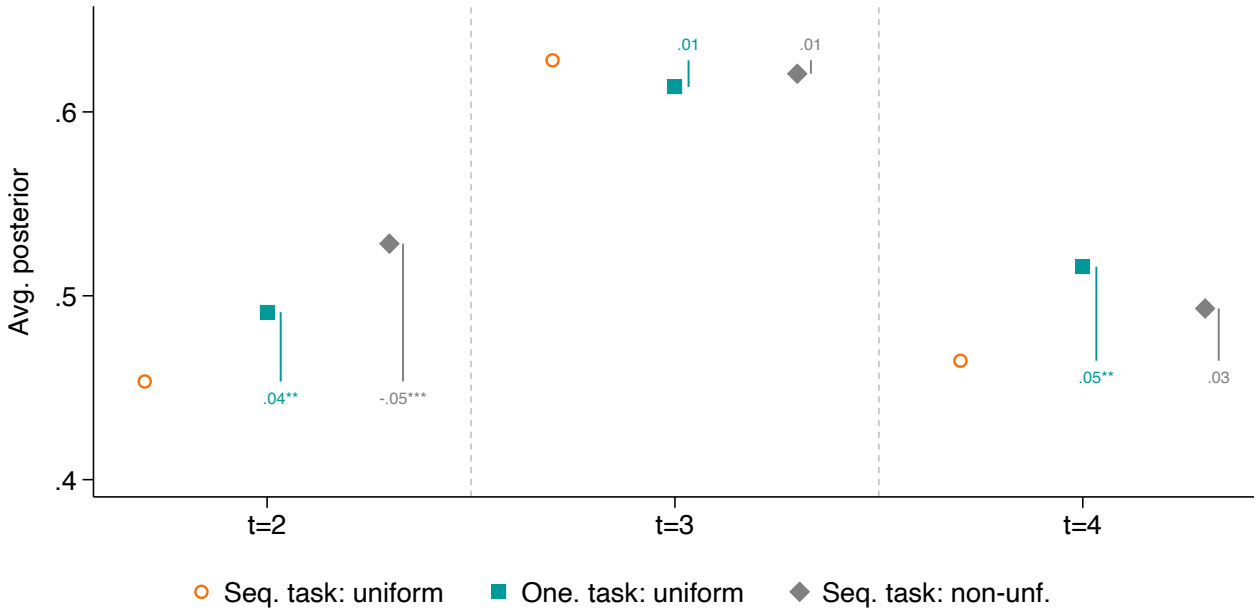


Figure A12: CDF: Posterior in sequential task, by prior
Sequence {1,0,1,1}
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

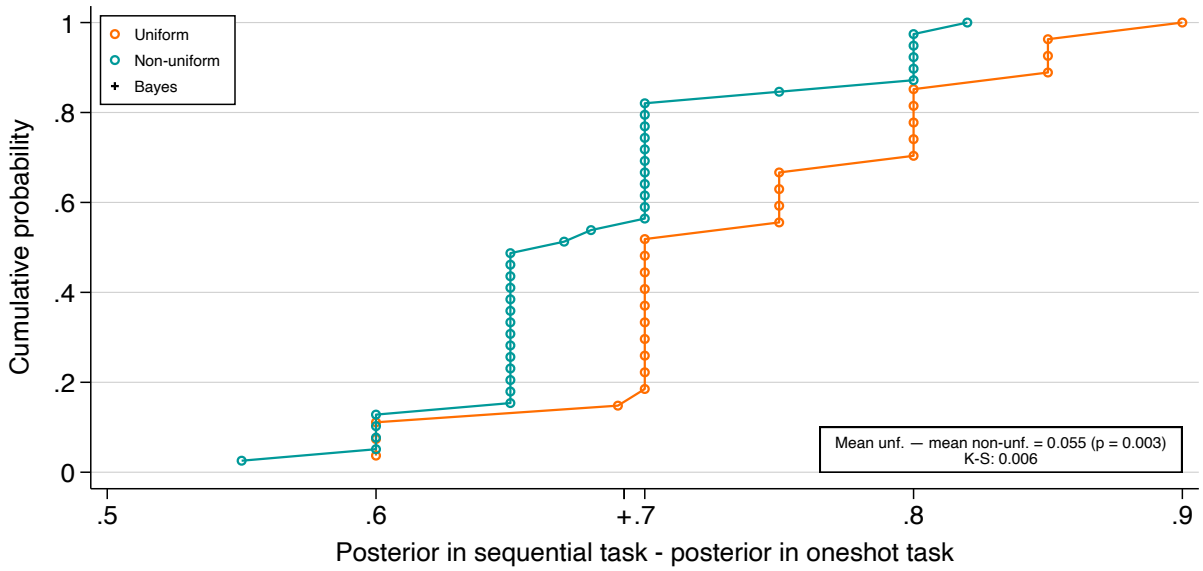


Figure A13: CDF: Posterior in sequential task, by prior
Sequence {1,1,0,1}
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

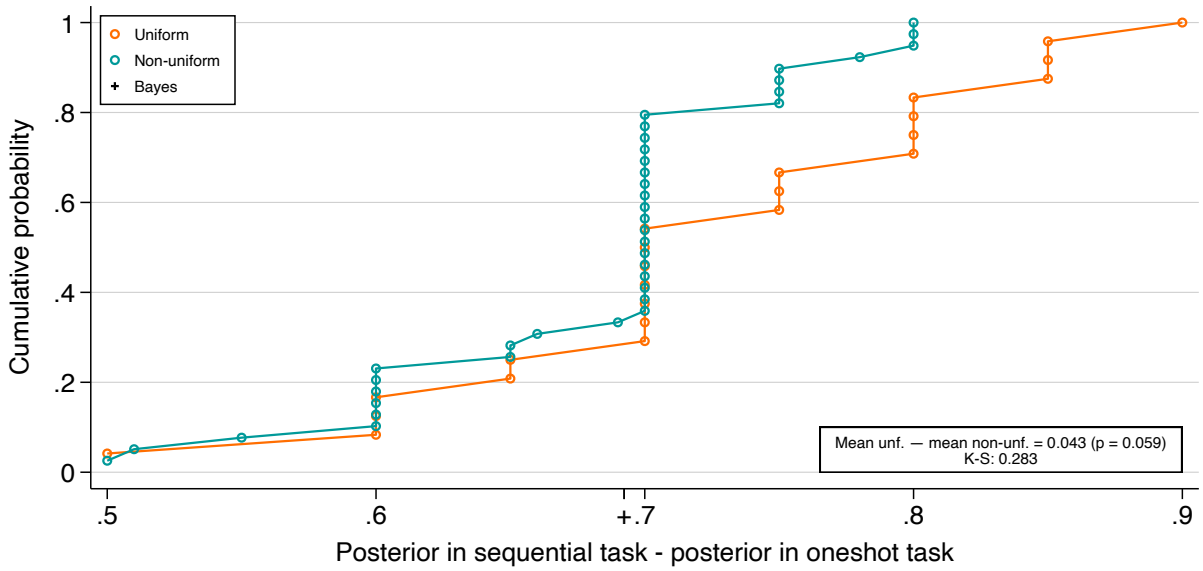


Figure A14: CDF: Posterior in sequential task, by prior
Sequence {1,1,1,0}
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

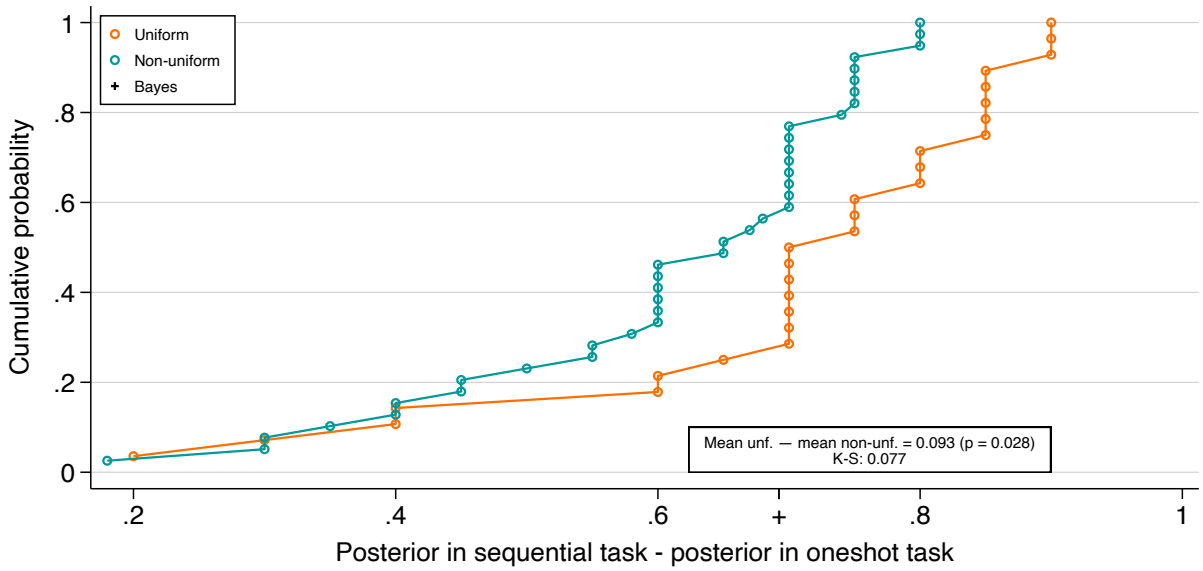


Figure A15: Avg. report by task and time
Sequence: {0,1,1,0}
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

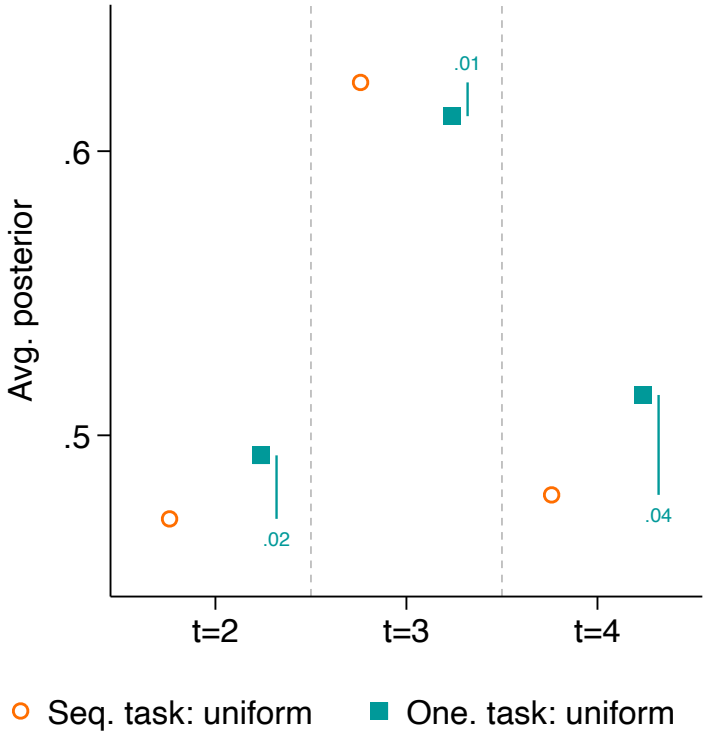


Figure A16: Avg. report by task and time
Sequence: {1,1,0,0}
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

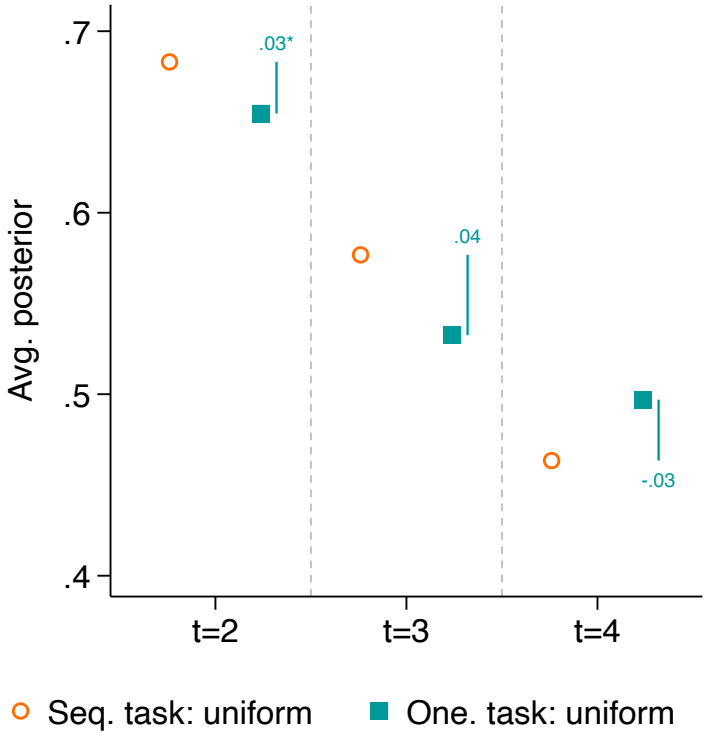


Figure A17: CDF: Posterior in sequential task, by prior
Sequence {1,0,1,1}
$q = 0.8$

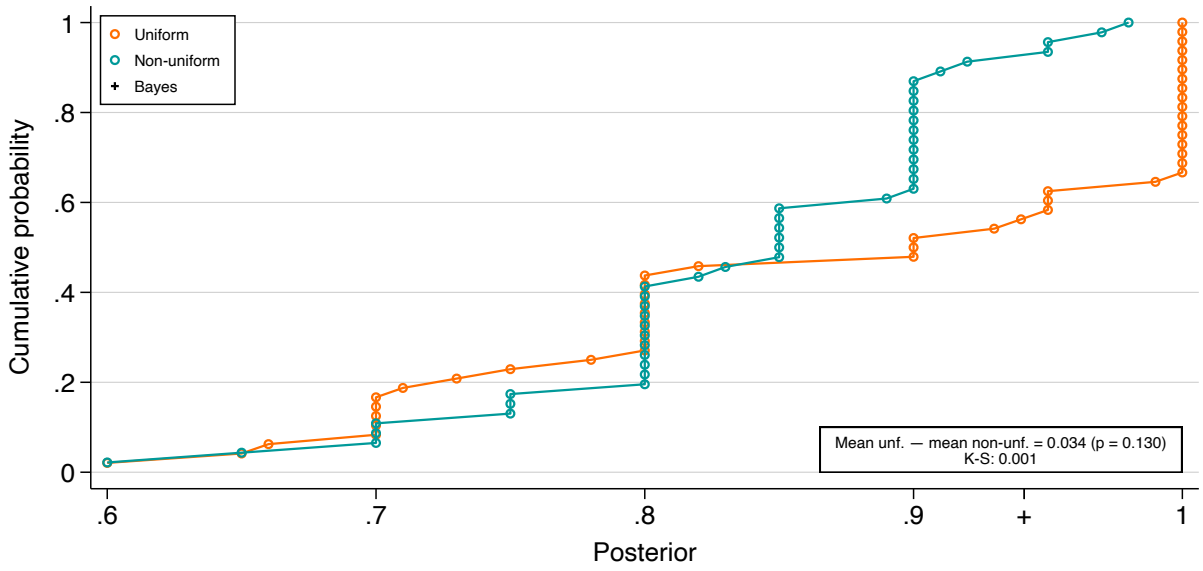


Figure A18: CDF: Posterior in sequential task, by prior
Sequence {1,1,0,1}
$q = 0.8$

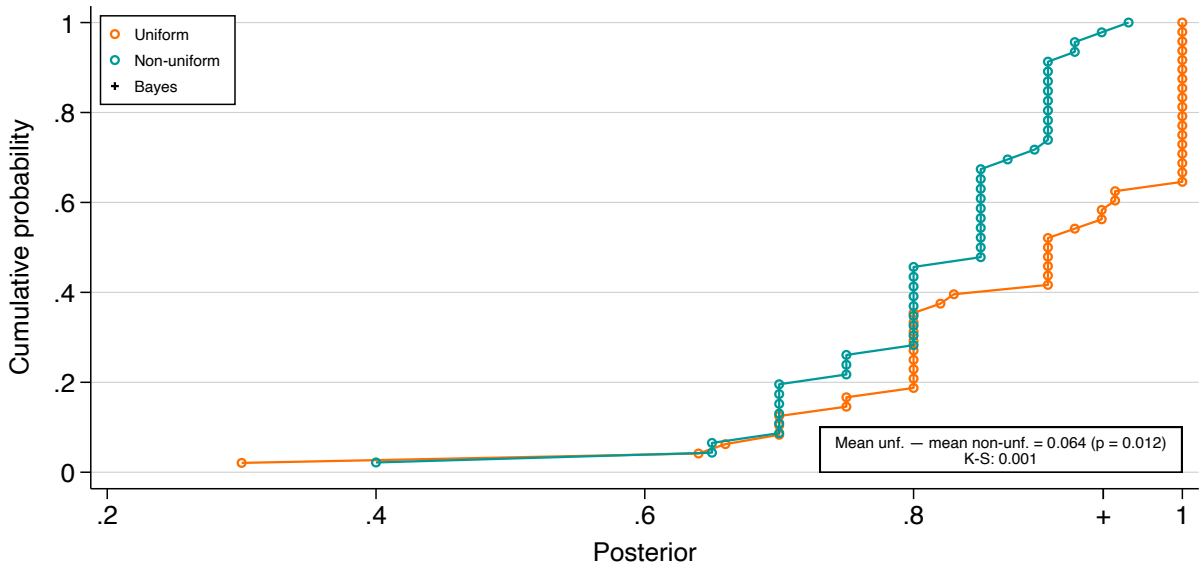


Figure A19: CDF: Posterior in sequential task, by prior
Sequence {1,1,0}
$q = 0.6$

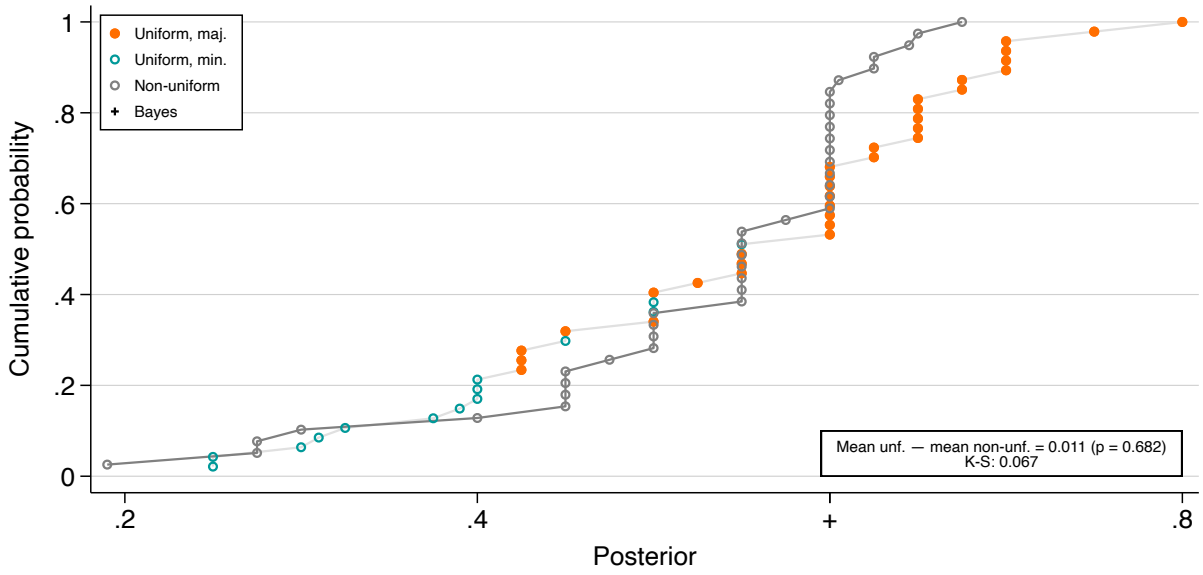
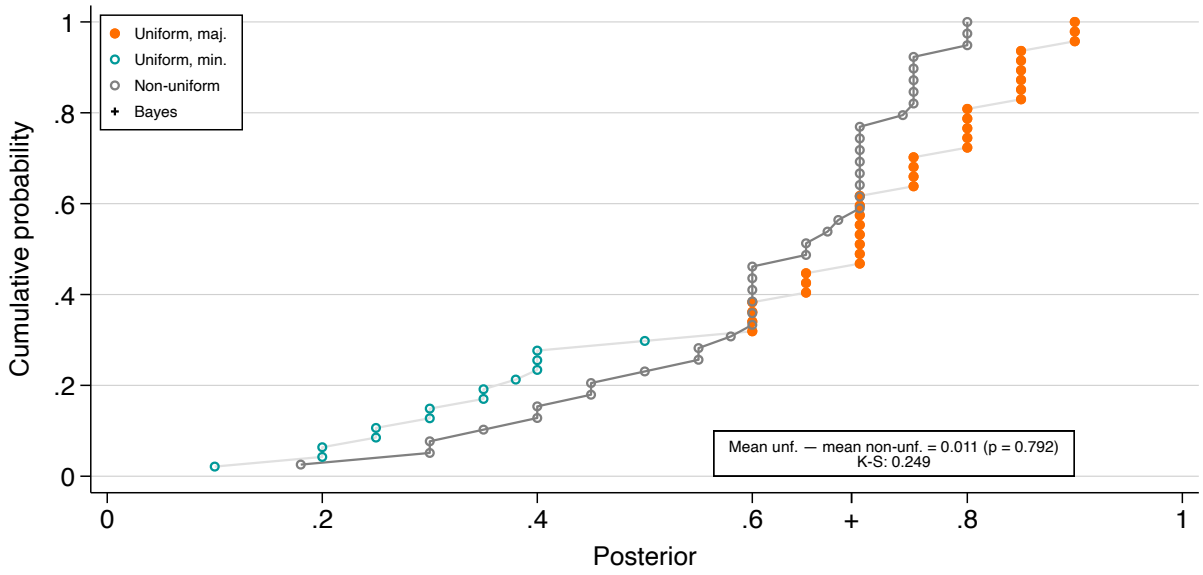


Figure A20: CDF: Posterior in sequential task, by prior
Sequence {1,1,1,0}
$q = 0.6$



Sources of prior sufficiency violations: aggregating signals

Figure A21: CDF: Posterior in sequential task
Sequences that reduce to one 1 signal
$p = 0.5$, $q = 0.6$; Maj. type

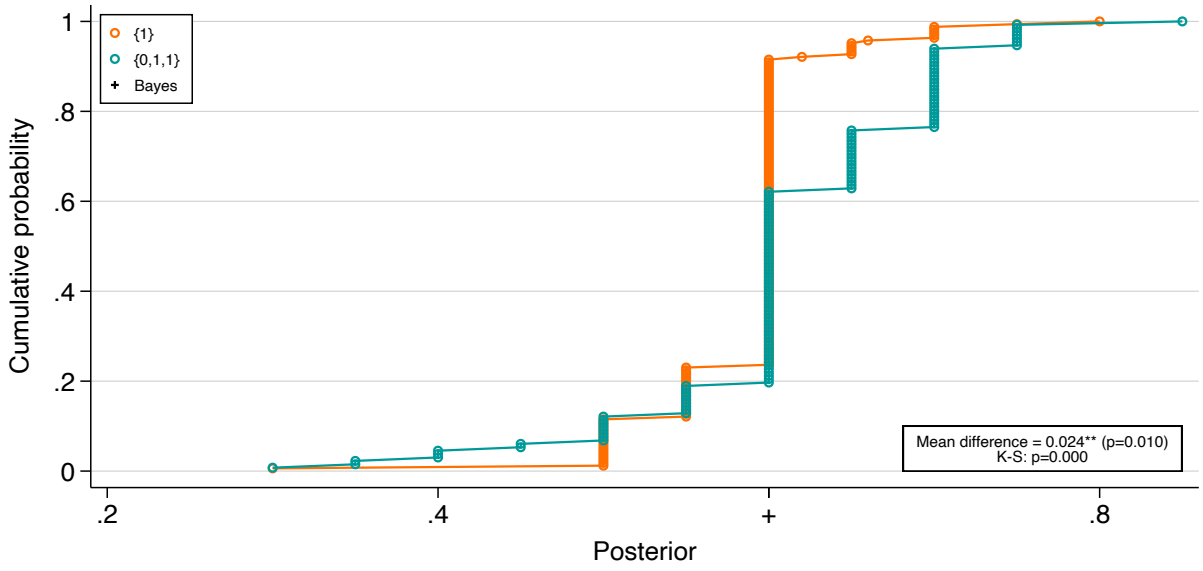


Figure A22: CDF: Posterior in one-shot task
Sequences that reduce to two 1 signals
$p = 0.5$, $q = 0.6$; Maj. type

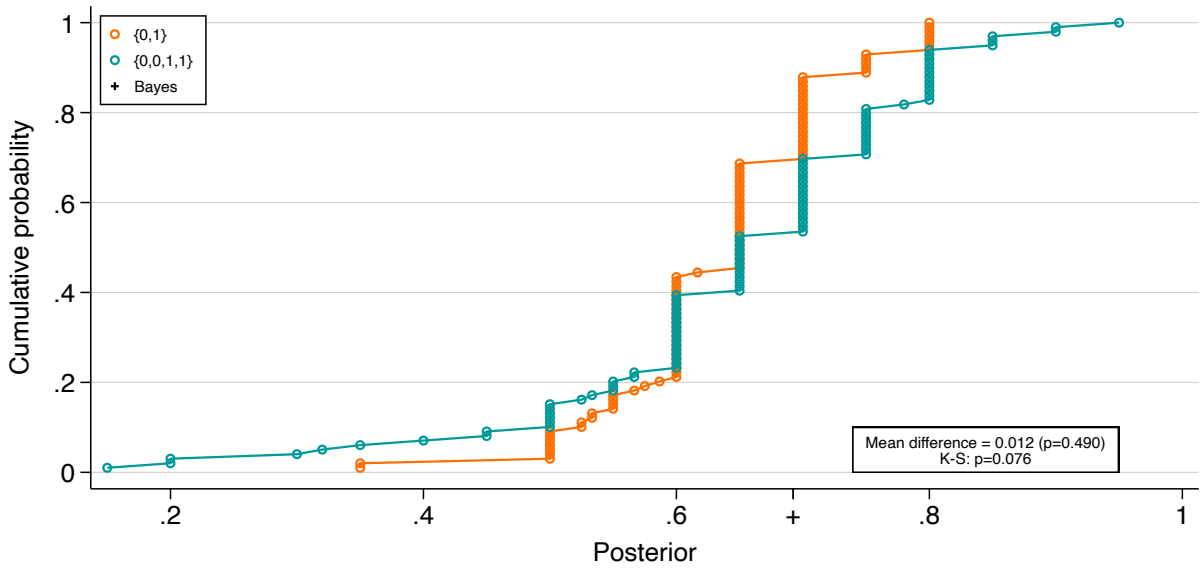
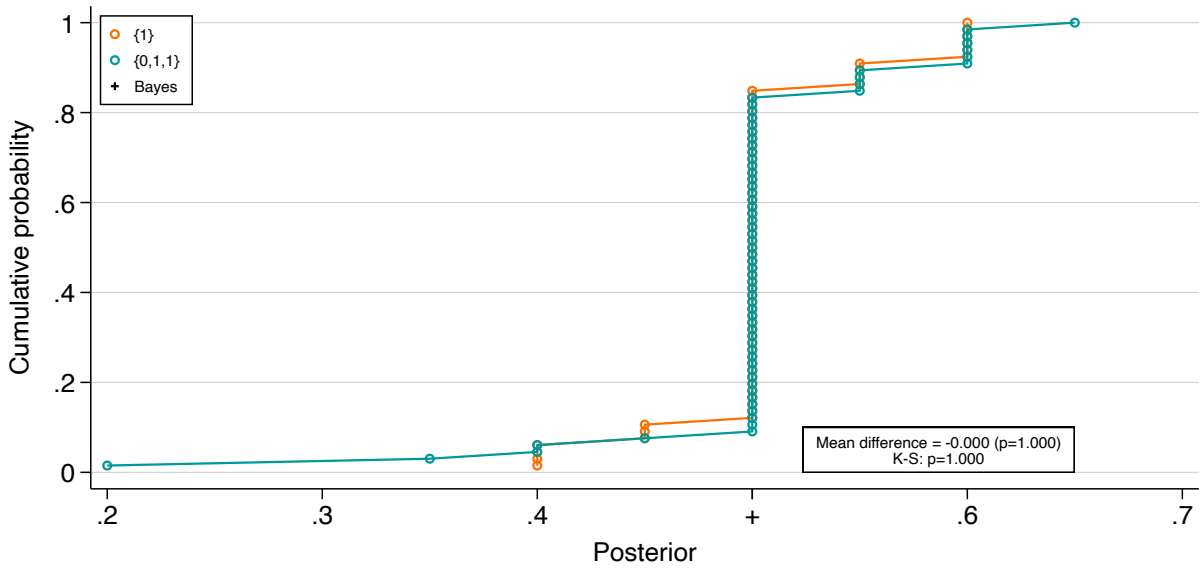


Figure A23: CDF: Posterior in sequential task
Uninformative sequences
$p = 0.5$, $q = 0.6$; Maj. type



Sources of prior sufficiency violations: updating in the non-uniform treatments

Figure A24: CDF: Posterior in sequential vs. one-shot task
Sequences {1,0,1,1} & {1,1,0,1}
$p = 0.8$, $q = 0.8$

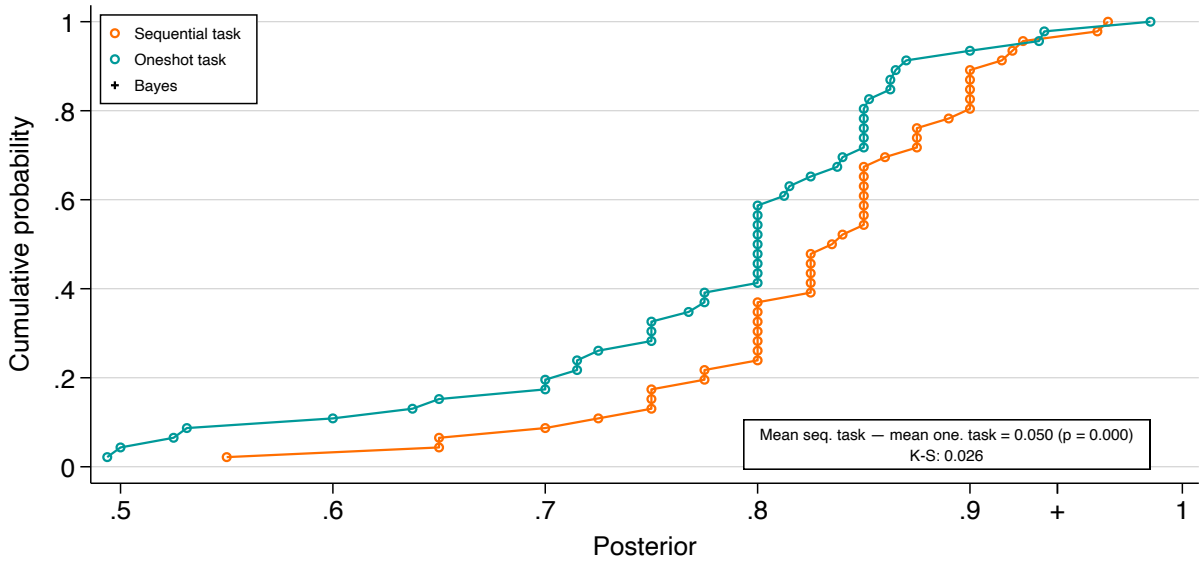


Figure A25: CDF: Posterior in sequential vs. one-shot task
Sequences {1,1,1,0}
$p = 0.6$, $q = 0.6$

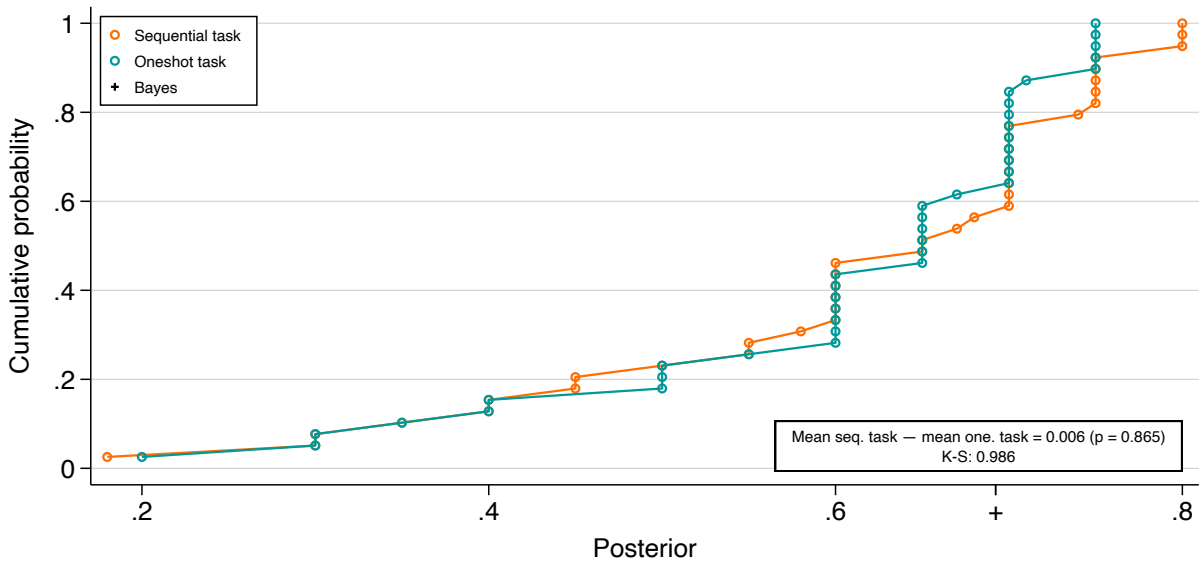


Figure A26: CDF: Posterior in sequential vs. one-shot task
Sequences {1,1,1,0}
$p = 0.8$, $q = 0.8$

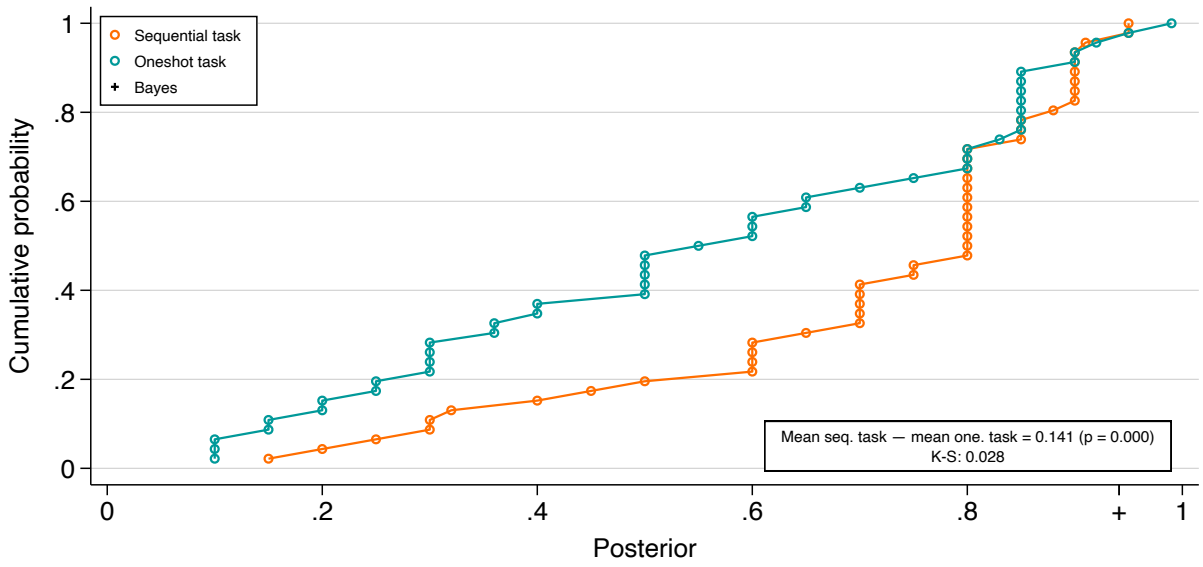


Figure A27: CDF: Posterior in sequential vs. one-shot task
Sequences of length 3 with one vs. two 1 signals
$p = 0.5$, $q = 0.6$

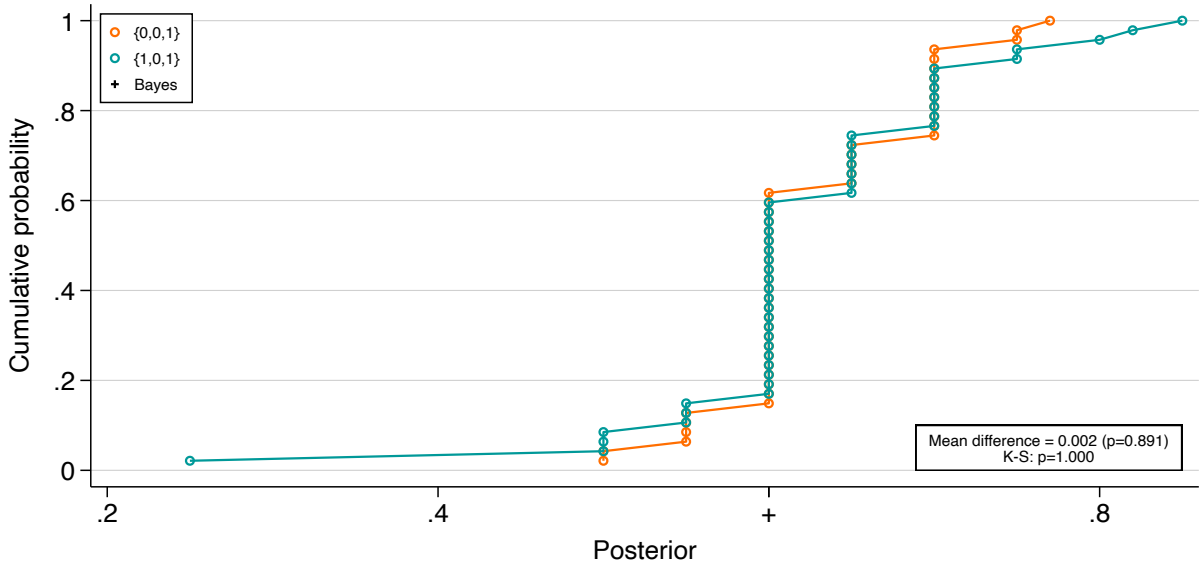


Figure A28: CDF: Posterior in sequential vs. one-shot task
Sequences of length 3 with one vs. two 1 signals
$p = 0.8$, $q = 0.8$

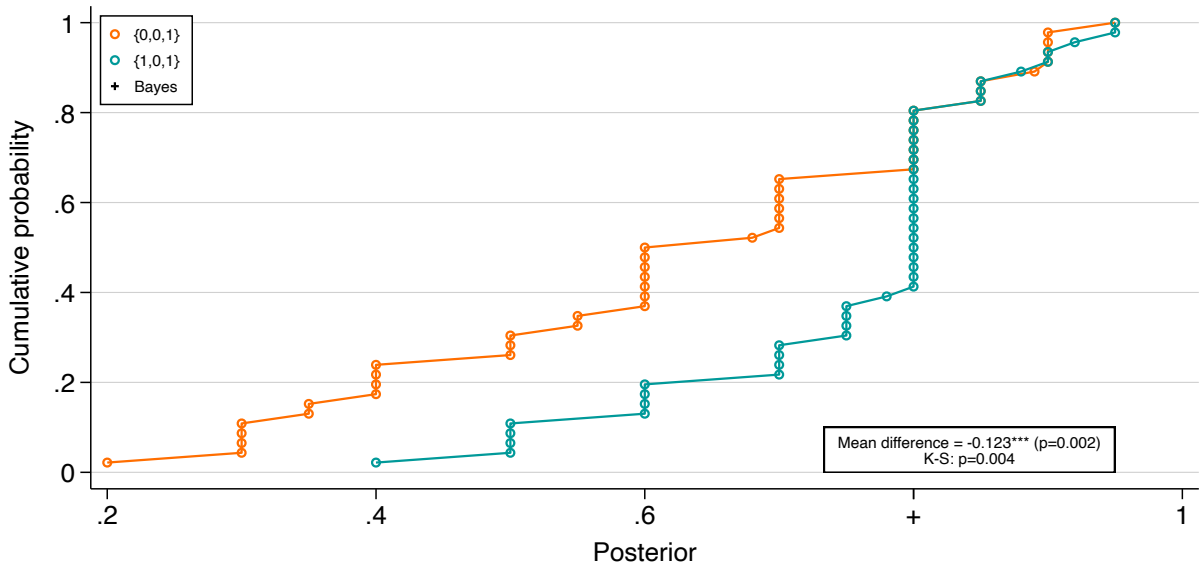


Figure A29: CDF: Posterior in sequential vs. one-shot task
Sequences {0,1,1} & {1,0,1}
$p = 0.6$, $q = 0.6$

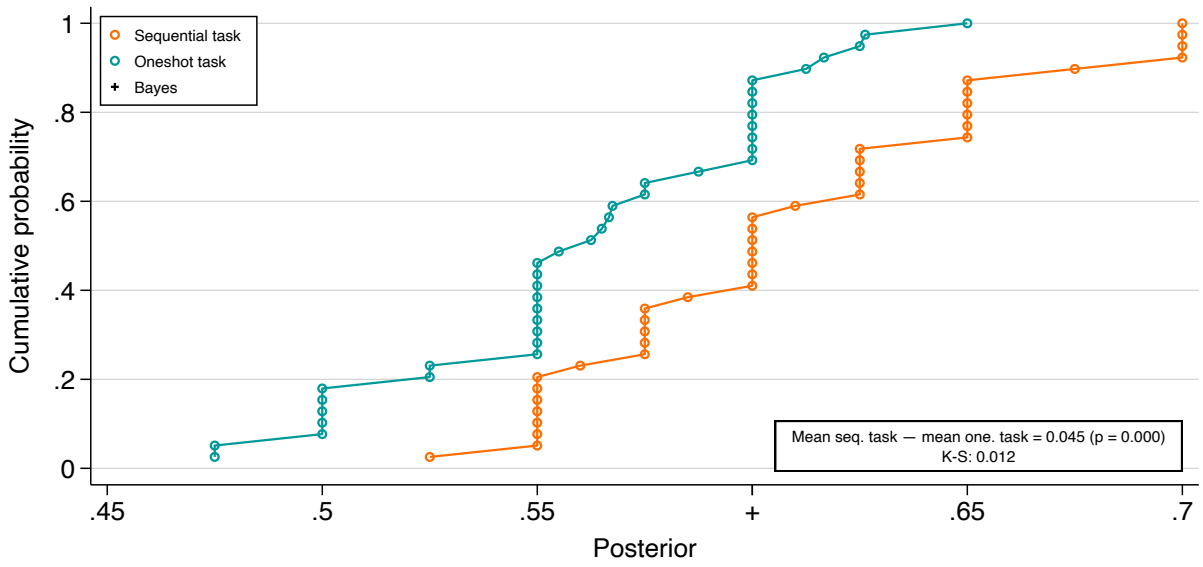


Figure A30: CDF: Posterior in sequential vs. one-shot task
Sequences {0,1,1} & {1,0,1}
$p = 0.8$, $q = 0.8$

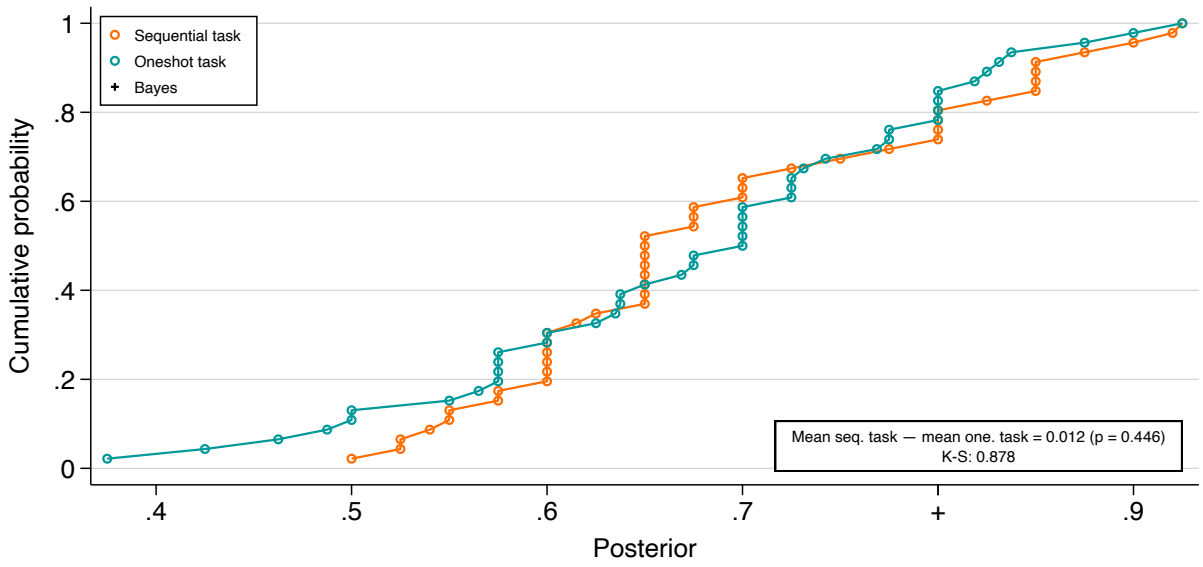


Figure A31: CDF: Posterior in sequential vs. one-shot task
Sequences {1,1,0}
$p = 0.6$, $q = 0.6$

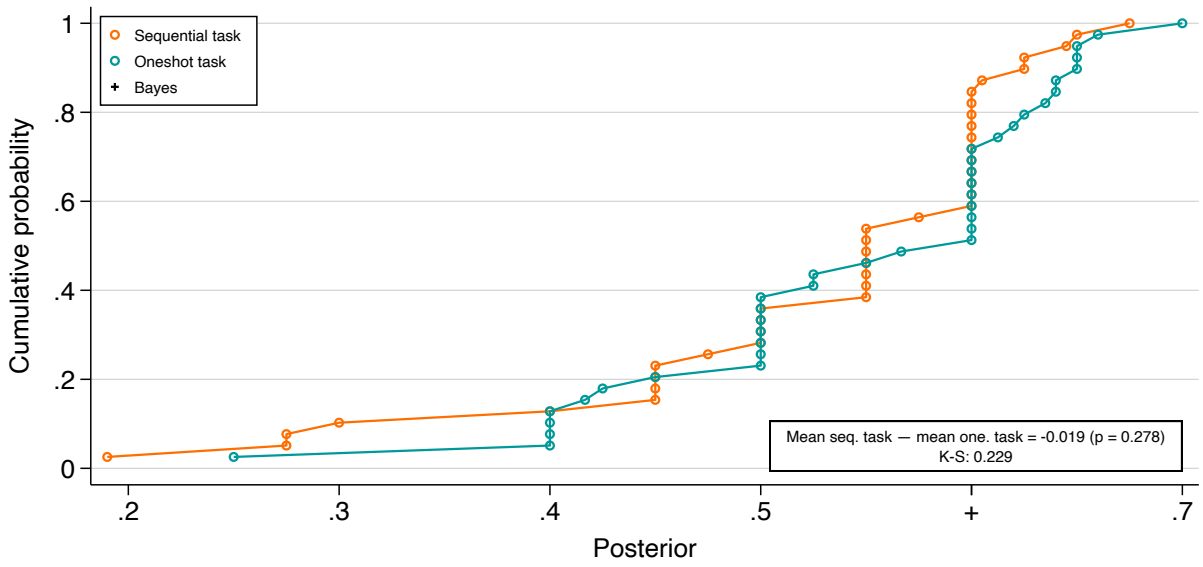
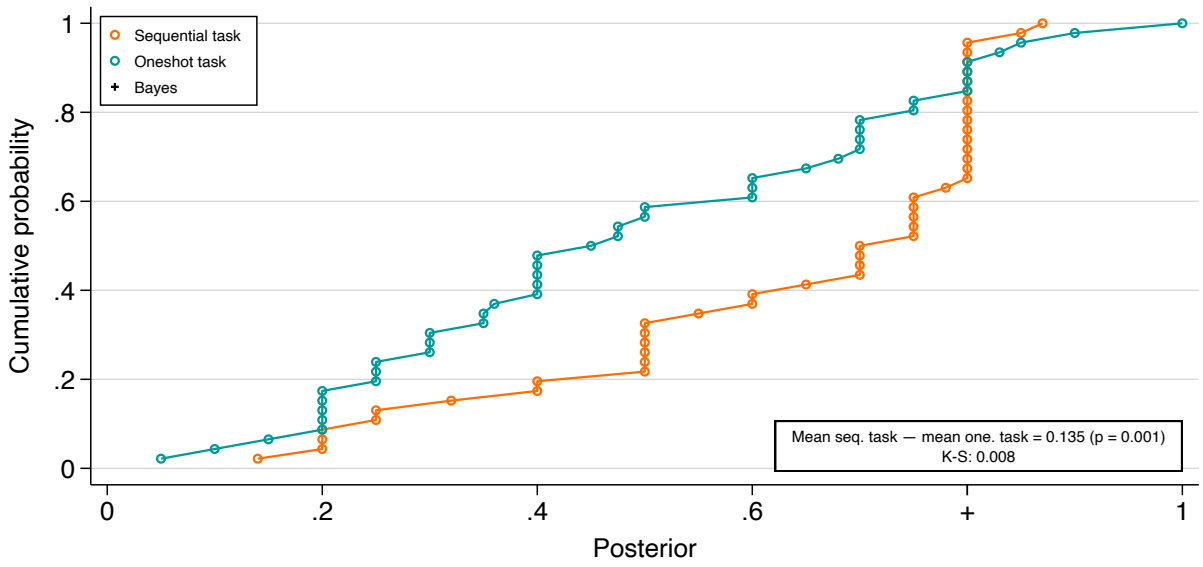


Figure A32: CDF: Posterior in sequential vs. one-shot task
Sequences {1,1,0}
$p = 0.8$, $q = 0.8$



B Tables

Table A1: Mean posterior by sequence:
$p = .5$, $q = .6$

s={0,1}				
Sequence	Mean posterior	{1,0}		{0,1,1,0}
{1,0}	0.459			
{0,1,1,0}	0.465	0.006		
{1,1,0,0}	0.446	-0.012		-0.018

s={1}				
Sequence	Mean posterior	{1}	{0,1,1}	{1,0,1}
{1}	0.606			
{0,1,1}	0.628	0.022**		
{1,0,1}	0.626	0.020	-0.002	
{1,1,0}	0.538	-0.068***	-0.090***	-0.088***

s={1,1}				
Sequence	Mean posterior	{1,1}	{1,0,1,1}	{1,1,0,1}
{1,1}	0.677			
{1,0,1,1}	0.716	0.039***		
{1,1,0,1}	0.727	0.050***	0.010	
{1,1,1,0}	0.620	-0.057	-0.096***	-0.107***

s={1,1,1}	
Sequence	Mean posterior
{1,1,1}	0.767

Notes: The first column lists the average posterior report for the sequence in the row. The remaining columns list the difference between the average report for the sequence in the row and the average report for the sequence in the column, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Each panel contains all sequences that reduce to the specified number of 0 signals and 1 signals. Standard errors are clustered at the individual level.

Table A2: Mean posterior by sequence:
$p = .5$, $q = .8$

s={0,1}				
Sequence	Mean posterior		{1,0}	
{1,0}	0.435			
{0,1,1,0}	0.463		0.028	

s={1}				
Sequence	Mean posterior	{1}	{0,1,1}	{1,0,1}
{1}	0.683			
{0,1,1}	0.729	0.045***		
{1,0,1}	0.696	0.013	-0.032*	
{1,1,0}	0.581	-0.103***	-0.148***	-0.116***

s={1,1}				
Sequence	Mean posterior	{1,1}	{1,0,1,1}	{1,1,0,1}
{1,1}	0.767			
{1,0,1,1}	0.871	0.104***		
{1,1,0,1}	0.878	0.111***	0.007	
{1,1,1,0}	0.728	-0.038	-0.143***	-0.150***

s={1,1,1}	
Sequence	Mean posterior
{1,1,1}	0.901

s={1,1,1,1}	
Sequence	Mean posterior
{1,1,1,1}	0.952

Notes: The first column lists the average posterior report for the sequence in the row. The remaining columns list the difference between the average report for the sequence in the row and the average report for the sequence in the column, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Each panel contains all sequences that reduce to the specified number of 0 signals and 1 signals. Standard errors are clustered at the individual level.

Table A3: Mean posterior by sequence:
$p = .5$, $q = .6$; Maj. type

s={0,1}				
Sequence	Mean posterior	{1,0}		{0,1,1,0}
{1,0}	0.492			
{0,1,1,0}	0.506	0.014		
{1,1,0,0}	0.502	0.010		-0.005

s={1}				
Sequence	Mean posterior	{1}	{0,1,1}	{1,0,1}
{1}	0.589			
{0,1,1}	0.620	0.031***		
{1,0,1}	0.624	0.036***	0.005	
{1,1,0}	0.603	0.014	-0.017	-0.021

s={1,1}				
Sequence	Mean posterior	{1,1}	{1,0,1,1}	{1,1,0,1}
{1,1}	0.667			
{1,0,1,1}	0.720	0.053***		
{1,1,0,1}	0.746	0.079***	0.027	
{1,1,1,0}	0.750	0.083***	0.030	0.004

s={1,1,1}	
Sequence	Mean posterior
{1,1,1}	0.784

Notes: The first column lists the average posterior report for the sequence in the row. The remaining columns list the difference between the average report for the sequence in the row and the average report for the sequence in the column, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Each panel contains all sequences that reduce to the specified number of 0 signals and 1 signals. Standard errors are clustered at the individual level.

Table A4: Mean posterior by sequence:
$p = .5$, $q = .8$; Maj. type

s={0,1}				
Sequence	Mean posterior		{1,0}	
{1,0}	0.483			
{0,1,1,0}	0.518		0.035	
s={1}				
Sequence	Mean posterior	{1}	{0,1,1}	{1,0,1}
{1}	0.660			
{0,1,1}	0.712	0.052***		
{1,0,1}	0.677	0.017	-0.036*	
{1,1,0}	0.664	0.004	-0.048*	-0.012
s={1,1}				
Sequence	Mean posterior	{1,1}	{1,0,1,1}	{1,1,0,1}
{1,1}	0.762			
{1,0,1,1}	0.890	0.128***		
{1,1,0,1}	0.893	0.131***	0.002	
{1,1,1,0}	0.871	0.109***	-0.019	-0.022
s={1,1,1}				
Sequence	Mean posterior			
{1,1,1}	0.927			
s={1,1,1,1}				
Sequence	Mean posterior			
{1,1,1,1}	0.973			

Notes: The first column lists the average posterior report for the sequence in the row. The remaining columns list the difference between the average report for the sequence in the row and the average report for the sequence in the column, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Each panel contains all sequences that reduce to the specified number of 0 signals and 1 signals. Standard errors are clustered at the individual level.

Table A5: Mean posterior by sequence:
$p = .5$, $q = .6$; Min. type

s={0,1}			
Sequence	Mean posterior	{1,0}	{0,1,1,0}
{1,0}	0.381		
{0,1,1,0}	0.367	-0.014	
{1,1,0,0}	0.316	-0.065	-0.051

s={1}				
Sequence	Mean posterior	{1}	{0,1,1}	{1,0,1}
{1}	0.647			
{0,1,1}	0.648	0.001		
{1,0,1}	0.630	-0.017	-0.018	
{1,1,0}	0.386	-0.261***	-0.262***	-0.244***

s={1,1}				
Sequence	Mean posterior	{1,1}	{1,0,1,1}	{1,1,0,1}
{1,1}	0.701			
{1,0,1,1}	0.708	0.007		
{1,1,0,1}	0.680	-0.021	-0.028	
{1,1,1,0}	0.313	-0.388***	-0.395***	-0.367***

s={1,1,1}	
Sequence	Mean posterior
{1,1,1}	0.726

Notes: The first column lists the average posterior report for the sequence in the row. The remaining columns list the difference between the average report for the sequence in the row and the average report for the sequence in the column, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Each panel contains all sequences that reduce to the specified number of 0 signals and 1 signals. Standard errors are clustered at the individual level.

Table A6: Mean posterior by sequence:
$p = .5$, $q = .8$; Min. type

s={0,1}				
Sequence	Mean posterior		{1,0}	
{1,0}	0.272			
{0,1,1,0}	0.279		0.007	

s={1}				
Sequence	Mean posterior	{1}	{0,1,1}	{1,0,1}
{1}	0.762			
{0,1,1}	0.783	0.020		
{1,0,1}	0.762	-0.001	-0.021	
{1,1,0}	0.299	-0.463***	-0.484***	-0.463***

s={1,1}				
Sequence	Mean posterior	{1,1}	{1,0,1,1}	{1,1,0,1}
{1,1}	0.783			
{1,0,1,1}	0.806	0.023		
{1,1,0,1}	0.829	0.046	0.023	
{1,1,1,0}	0.249	-0.534***	-0.557***	-0.580***

s={1,1,1}	
Sequence	Mean posterior
{1,1,1}	0.813

s={1,1,1,1}	
Sequence	Mean posterior
{1,1,1,1}	0.882

Notes: The first column lists the average posterior report for the sequence in the row. The remaining columns list the difference between the average report for the sequence in the row and the average report for the sequence in the column, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Each panel contains all sequences that reduce to the specified number of 0 signals and 1 signals. Standard errors are clustered at the individual level.

Table A7: Mean posterior report
$q = 0.6$

Signal composition	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.659	0.000	0.659***
{0,1,1}	0.568	0.554	0.013
{0,0,1,1}	0.466	0.490	-0.023**
{0,1,1,1}	0.658	0.619	0.039***
N Participants	39	39	39
N Observations	468	351	351

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A8: Mean posterior report
$q = 0.8$

Signal composition	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.809	0.000	0.809***
{0,1,1}	0.670	0.617	0.053***
{0,0,1,1}	0.472	0.390	0.083*
{0,1,1,1}	0.784	0.703	0.080***
N Participants	46	46	46
N Observations	437	299	299

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A9: Mean posterior report
$q = 0.6$

Sequence	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.659	0.000	0.659***
{0,1,1} & {1,0,1}	0.608	0.562	0.045***
{1,0,1,1} & {1,1,0,1}	0.683	0.627	0.056***
{1,1,0}	0.527	0.546	-0.019
{1,1,1}	0.731	0.696	0.035***
{1,1,1,0}	0.609	0.603	0.006
{0,0,0,1,1}	0.710	0.603	0.107***
{0,0,1,1,1}	0.531	0.569	-0.038**
N Participants	39	39	39
N Observations	565	448	448

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A10: Mean posterior report
$q = 0.8$

Sequence	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.809	0.000	0.809***
{0,1,1} & {1,0,1}	0.691	0.679	0.012
{1,0,1,1} & {1,1,0,1}	0.826	0.776	0.050***
{1,1,0}	0.626	0.492	0.135***
{1,1,1}	0.868	0.823	0.046**
{0,1,1,1}	0.755	0.708	0.046
{1,1,1,0}	0.699	0.558	0.141***
{1,1,1,1}	0.922	0.849	0.073***
{0,0,0,0,1}	0.911	0.834	0.077**
{0,0,0,1,1}	0.784	0.663	0.121**
{0,0,1,1,1}	0.585	0.580	0.004
{0,1,1,1,1}	0.892	0.842	0.049***
N Participants	46	46	46
N Observations	759	621	621

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A11: Mean posterior report
2 states, $q = 0.8$, restrict to obs. from 2 state uniform
with exact match to oneshot

Sequence	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.809	0.000	0.809***
{0,1,1} & {1,0,1}	0.691	0.679	0.012
{1,0,1,1} & {1,1,0,1}	0.826	0.776	0.050***
{1,1,0}	0.626	0.492	0.135***
{1,1,1}	0.868	0.823	0.046**
{0,1,1,1}	0.755	0.708	0.046
{1,1,1,0}	0.699	0.558	0.141***
{1,1,1,1}	0.922	0.849	0.073***
{0,0,0,0,1}	0.911	0.834	0.077**
{0,0,0,1,1}	0.784	0.663	0.121**
{0,0,1,1,1}	0.585	0.580	0.004
{0,1,1,1,1}	0.892	0.842	0.049***
N Participants	46	46	46
N Observations	759	621	621

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A12: Mean posterior report
$q = 0.8$, restrict to obs. where prior less than 1

Sequence	Seq. task Unf.	One. task Unf.	Seq. task Non-unf.	Seq. Unf — One. Unf.	Seq. Unf — Seq. Non-unf.
{1,0}	0.435	0.458	0.521	-0.023	-0.086$^{**}_{*}$
{1,1}	0.767	0.772	0.809	-0.006	-0.042$^{**}_{*}$
{0,1,1} & {1,0,1}	0.712	0.671	0.691	0.041***	0.021
{1,0,1,1} & {1,1,0,1}	0.875	0.738	0.826	0.136***	0.049$^{**}_{***}$
{1,1,0}	0.581	0.506	0.626	0.075*	-0.046***
{1,1,1}	0.903	0.838	0.868	0.065***	0.035$^{*}_{*}$
{1,1,1,0}	0.649	0.539	0.699	0.109*	-0.050***
N Participants	48	48	46	48	46
N Observations	606	606	598	606	585

Notes: Columns 1-3 list the mean posterior report for the treatment in the column and sequence in the row. Columns 4-6 list the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Red stars correspond to p-value for difference-in-difference of (Seq. Unf. - One. Unf.) and (Seq Unf. - Seq. Non-unf.) Standard errors are clustered at the individual level.

Table A13: Mean posterior report
2 states, $q = 0.6$, restrict to obs. from 2 state uniform
with exact match to oneshot

Signal composition	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.659	0.000	0.659***
{0,1,1}	0.568	0.554	0.013
{0,0,1,1}	0.466	0.490	-0.023**
{0,1,1,1}	0.658	0.619	0.039***
N Participants	39	39	39
N Observations	468	351	351

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A14: Mean posterior report
2 states, $q = 0.8$, restrict to obs. from 2 state uniform
with exact match to oneshot

Signal composition	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.809	0.000	0.809***
{0,1,1}	0.670	0.617	0.053***
{0,0,1,1}	0.472	0.390	0.083*
{0,1,1,1}	0.784	0.703	0.080***
N Participants	46	46	46
N Observations	437	299	299

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A15: Mean posterior report
2 states, $q = 0.6$, restrict to obs. from 2 state uniform
with exact match to oneshot

Sequence	Seq. task Non-unf.	One. task Non-unf.	Seq. Non-unf − One. Non-unf.
{1,1}	0.659	0.000	0.659***
{0,1,1} & {1,0,1}	0.608	0.562	0.045***
{1,0,1,1} & {1,1,0,1}	0.683	0.627	0.056***
{1,1,0}	0.527	0.546	-0.019
{1,1,1}	0.731	0.696	0.035***
{1,1,1,0}	0.609	0.603	0.006
{0,0,0,1,1}	0.710	0.603	0.107***
{0,0,1,1,1}	0.531	0.569	-0.038**
N Participants	39	39	39
N Observations	565	448	448

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A16: Mean posterior report
$q = 0.6$
Maj. type

Sequence	Seq. task Unf.	One. task Unf.	Seq. Unf − One. Unf.
{0,1}	0.488	0.489	-0.001
{1,0}	0.495	0.516	-0.021
{1,1}	0.667	0.635	0.031**
{0,1,1}	0.620	0.592	0.027**
{1,0,1}	0.624	0.568	0.057***
{1,1,0}	0.603	0.516	0.087***
{1,1,1}	0.784	0.728	0.056***
{0,1,1,0}	0.494	0.480	0.014
{1,0,1,1}	0.720	0.678	0.042*
{1,1,0,0}	0.502	0.514	-0.012
{1,1,0,1}	0.746	0.647	0.100***
{1,1,1,0}	0.750	0.619	0.131***
N Participants	33	33	33
N Observations	495	495	495

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A17: Mean posterior report
$q = 0.6$
Maj. type

Sequence	Seq. task Unf.	One. task Unf.	Seq. Unf – One. Unf.
{1,0}	0.504	0.513	-0.010
{1,1}	0.667	0.635	0.031**
{0,1,1} & {1,0,1}	0.622	0.580	0.042***
{1,0,1,1} & {1,1,0,1}	0.733	0.662	0.071***
{1,1,0}	0.603	0.516	0.087***
{1,1,1}	0.784	0.728	0.056***
{1,1,1,0}	0.750	0.619	0.131***
N Participants	33	33	33
N Observations	429	429	429

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A18: Mean posterior report
$q = 0.6$
Min. type

Sequence	Seq. task Unf.	One. task Unf.	Seq. Unf − One. Unf.
{0,1}	0.621	0.505	0.115***
{1,0}	0.383	0.458	-0.074
{1,1}	0.701	0.708	-0.007
{0,1,1}	0.648	0.664	-0.016
{1,0,1}	0.630	0.555	0.075
{1,1,0}	0.386	0.534	-0.148***
{1,1,1}	0.726	0.753	-0.027
{0,1,1,0}	0.633	0.494	0.139***
{1,0,1,1}	0.708	0.713	-0.005
{1,1,0,0}	0.316	0.320	-0.004
{1,1,0,1}	0.680	0.546	0.134***
{1,1,1,0}	0.313	0.507	-0.194***
N Participants	14	14	14
N Observations	210	210	210

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A19: Mean posterior report
$q = 0.8$
Maj. type

Sequence	Seq. task Unf.	One. task Unf.	Seq. Unf − One. Unf.
{1,0}	0.529	0.534	-0.005
{1,1}	0.762	0.757	0.005
{0,1,1} & {1,0,1}	0.695	0.656	0.039**
{1,0,1,1} & {1,1,0,1}	0.891	0.760	0.132***
{1,1,0}	0.664	0.527	0.137***
{1,1,1}	0.927	0.829	0.098***
{1,1,1,0}	0.871	0.599	0.272***
N Participants	37	37	37
N Observations	481	481	481

Notes: Columns 1 and 2 list the mean posterior report for the treatment in the column and sequence in the row. Column 3 lists the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Standard errors are clustered at the individual level.

Table A20: Mean posterior report
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

Sequence	Seq. task Unf.	One. task Unf.	Seq. task Non-unf.	Seq. Unf − One. Unf.	Seq. Unf − Seq. Non-unf.
{1,0}	0.482	0.503	0.472	-0.021	0.010*
{1,1}	0.683	0.655	0.659	0.028*	0.024*
{0,1,1} & {1,0,1}	0.623	0.598	0.608	0.026***	0.015
{1,0,1,1} & {1,1,0,1}	0.732	0.664	0.683	0.068***	0.049***
{1,1,0}	0.577	0.533	0.527	0.044	0.049*
{1,1,1}	0.790	0.750	0.731	0.040**	0.059***
{1,1,1,0}	0.702	0.623	0.609	0.079*	0.093**
N Participants	39	39	39	39	39
N Observations	363	363	507	363	363

Notes: Columns 1-3 list the mean posterior report for the treatment in the column and sequence in the row. Columns 4-6 list the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Red stars correspond to p-value for difference-in-difference of (Seq. Unf. - One. Unf.) and (Seq Unf. - Seq. Non-unf.) Standard errors are clustered at the individual level.

Table A21: Mean posterior report
$q = 0.6$, restrict to obs. from 2 state uniform with Bayesian posterior after 1st signal

Sequence	Seq. task Unf.	One. task Unf.	Seq. task Non-unf.	Seq. Unf — One. Unf.	Seq. Unf — Seq. Non-unf.
{0,1}	0.529	0.507	0.000	0.022	0.529*** ***
{1,0}	0.498	0.517	0.472	-0.019	0.026*
{1,1}	0.683	0.655	0.659	0.028*	0.024*
{0,1,1}	0.624	0.612	0.621	0.012	0.003
{1,0,1}	0.622	0.581	0.595	0.041***	0.027*
{1,1,0}	0.577	0.533	0.527	0.044	0.049*
{1,1,1}	0.790	0.750	0.731	0.040**	0.059***
{0,1,1,0}	0.521	0.486	0.507	0.035	0.014
{1,0,1,1}	0.739	0.692	0.684	0.047***	0.055***
{1,1,0,0}	0.463	0.497	0.439	-0.034	0.024*
{1,1,0,1}	0.725	0.632	0.682	0.093***	0.043**
{1,1,1,0}	0.702	0.623	0.609	0.079*	0.093**
N Participants	39	39	39	39	39
N Observations	426	426	585	426	392

Notes: Columns 1-3 list the mean posterior report for the treatment in the column and sequence in the row. Columns 4-6 list the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Red stars correspond to p-value for difference-in-difference of (Seq. Unf. - One. Unf.) and (Seq Unf. - Seq. Non-unf.) Standard errors are clustered at the individual level.

Table A22: Mean posterior report
$q = 0.8$

Sequence	Seq. task Unf.	One. task Unf.	Seq. task Non-unf.	Seq. Unf — One. Unf.	Seq. Unf — Seq. Non-unf.
{1,0}	0.509	0.535	0.521	-0.025	-0.012
{1,1}	0.767	0.772	0.809	-0.006	-0.042**
{0,1,1} & {1,0,1}	0.712	0.671	0.691	0.041***	0.021
{1,0,1,1} & {1,1,0,1}	0.875	0.738	0.826	0.136***	0.049***
{1,1,0}	0.581	0.506	0.626	0.075*	-0.046***
{1,1,1}	0.901	0.836	0.868	0.065***	0.033*
{1,1,1,0}	0.728	0.564	0.699	0.164***	0.029**
N Participants	48	48	46	48	46
N Observations	624	624	598	624	598

Notes: Columns 1-3 list the mean posterior report for the treatment in the column and sequence in the row. Columns 4-6 list the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Red stars correspond to p-value for difference-in-difference of (Seq. Unf. - One. Unf.) and (Seq Unf. - Seq. Non-unf.) Standard errors are clustered at the individual level.

Table A23: Mean posterior report
$q = 0.6$

Sequence	Seq. task Unf.	One. task Unf.	Seq. task Non-unf.	Seq. Unf — One. Unf.	Seq. Unf — Seq. Non-unf.
{0,1}	0.453	0.491	0.000	-0.038**	0.453*** ***
{1,0}	0.465	0.500	0.472	-0.035	-0.007
{1,1}	0.677	0.657	0.659	0.020	0.018
{0,1,1}	0.628	0.614	0.621	0.014	0.007
{1,0,1}	0.626	0.564	0.595	0.062***	0.031*
{1,1,0}	0.538	0.521	0.527	0.017	0.011
{1,1,1}	0.767	0.735	0.731	0.031**	0.036**
{0,1,1,0}	0.535	0.484	0.507	0.051**	0.028
{1,0,1,1}	0.716	0.688	0.684	0.028	0.032
{1,1,0,0}	0.446	0.456	0.439	-0.010	0.007
{1,1,0,1}	0.727	0.617	0.682	0.110***	0.045** ***
{1,1,1,0}	0.620	0.586	0.609	0.034	0.011
N Participants	47	47	39	47	39
N Observations	705	705	585	705	551

Notes: Columns 1-3 list the mean posterior report for the treatment in the column and sequence in the row. Columns 4-6 list the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Red stars correspond to p-value for difference-in-difference of (Seq. Unf. - One. Unf.) and (Seq Unf. - Seq. Non-unf.) Standard errors are clustered at the individual level.

Table A24: Mean posterior report
$q = 0.6$

Sequence	Seq. task Unf.	One. task Unf.	Seq. task Non-unf.	Seq. Unf — One. Unf.	Seq. Unf — Seq. Non-unf.
{1,0}	0.508	0.505	0.472	0.003	0.037**
{1,1}	0.677	0.657	0.659	0.020	0.018
{0,1,1} & {1,0,1}	0.627	0.589	0.608	0.038***	0.019
{1,0,1,1} & {1,1,0,1}	0.721	0.653	0.683	0.069***	0.039**
{1,1,0}	0.538	0.521	0.527	0.017	0.011
{1,1,1}	0.767	0.735	0.731	0.031**	0.036**
{1,1,1,0}	0.620	0.586	0.609	0.034	0.011
N Participants	47	47	39	47	39
N Observations	611	611	507	611	507

Notes: Columns 1-3 list the mean posterior report for the treatment in the column and sequence in the row. Columns 4-6 list the difference between the indicated treatments, with stars indicating significance: *** p<0.01, ** p<0.05, * p<0.10. Red stars correspond to p-value for difference-in-difference of (Seq. Unf. - One. Unf.) and (Seq Unf. - Seq. Non-unf.) Standard errors are clustered at the individual level.

Table A25: Signal sequences by treatment

Order 1	**Order 2**
$p = 0.5, q = 0.6$	
$\{1,1,1,1\}$	$\{0,0,0,0\}$
$\{0,0,1,1\}$	$\{1,0,0,1\}$
$\{0,1,0,0\}$	$\{1,1,1,0\}$
$\{1,1,0,1\}$	$\{0,0,1,0\}$
$\{0,1,1,0\}$	$\{1,0,1,1\}$
$\{0,0,0,1\}$	$\{1,1,0,0\}$
$p = 0.5, q = 0.8$	
$\{0,0,0,0\}$	$\{1,1,1,1\}$
$\{0,1,0,0\}$	$\{1,0,0,1\}$
$\{1,1,0,1\}$	$\{1,1,1,0\}$
$\{0,1,1,0\}$	$\{0,0,1,0\}$
$\{0,0,0,1\}$	$\{1,0,1,1\}$
$\{1,1,1,1\}$	$\{0,0,0,0\}$
$p = 0.6, q = 0.6$	
$\{1,1,1,1\}$	$\{0,0,0,0\}$
$\{1,0,0,0\}$	$\{1,0,1,0\}$
$\{0,1,1,0\}$	$\{1,1,0,1\}$
$\{1,0,1,1\}$	$\{0,0,1,1\}$
$\{0,0,1,0\}$	$\{0,1,1,1\}$
$\{1,1,0,0\}$	$\{1,0,0,0\}$
$p = 0.8, q = 0.8$	
$\{1,1,1,1\}$	$\{1,1,1,1\}$
$\{0,1,1,0\}$	$\{0,0,1,0\}$
$\{1,0,1,1\}$	$\{1,0,1,1\}$
$\{0,0,0,0\}$	$\{1,1,0,0\}$
$\{1,1,0,1\}$	$\{0,1,1,1\}$
$\{1,1,1,1\}$	$\{1,1,1,1\}$

Notes: Each panel shows the signal sequences for each order in the indicated treatment.

C Estimation of Grether models

C.1 Endogeneity of pooled OLS and instrumental variables

Suppose that we estimate the following regression, pooling over i and t:

$$\pi_{it} = \beta_i \lambda_{it} + \delta_i \pi_{i,t-1} + u_{it}$$

where π_{it} is the log posterior odds, $\pi_{i,t-1}$ is the log prior odds, λ_{it} is the log-likelihood ratio, and u_{it} is an i.i.d. error term. Define $\theta_i = [\delta_i \ \beta_i]'$ and $X_{it} = [\pi_{it} \ \lambda_{it}]'$. Denote the OLS estimate $\hat{\theta}$.

$$\begin{aligned}
\hat{\theta} &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}X_{it}'\right]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\pi_{it} \\
&= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}X_{it}'\right]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\left(X_{it}'\theta_i + u_{it}\right) \\
&= \underbrace{\left[\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}X_{it}'\right]^{-1}}_{A} \left(\underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}X_{it}'\theta_i}_{B} + \underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}u_{it}}_{C} \right)
\end{aligned}$$

$$A \cdot B = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \begin{bmatrix} \pi_{it}^2 & \pi_{it}s_{it} \\ \pi_{it}s_{it} & s_{it}^2 \end{bmatrix}\right)^{-1} \left(\underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T} \begin{bmatrix} \pi_{it}^2 & \pi_{it}s_{it} \\ \pi_{it}s_{it} & s_{it}^2 \end{bmatrix}}_{K} \begin{bmatrix} \delta_i \\ \beta_i \end{bmatrix} \right)$$

Given the assumption that u_{it} is uncorrelated with X_{it}, C is 0 in expectation. Since π_{it} is correlated with θ_i, we cannot pull K out. As a result, A and K do not cancel, and so OLS does not recover $\mathbb{E}(\delta_i)$ or $\mathbb{E}(\beta_i)$.

Now, suppose that we instrument π_{it} with some variable z_{it}. Denote the first stage

predicted value $\hat{\pi}_{it}$. Then we have:

$$A \cdot B = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \begin{bmatrix} \hat{\pi}_{it}^2 & \hat{\pi}_{it} s_{it} \\ \hat{\pi}_{it} s_{it} & s_{it}^2 \end{bmatrix} \right)^{-1} \left(\underbrace{\sum_{i=1}^{N} \sum_{t=1}^{T} \begin{bmatrix} \hat{\pi}_{it}^2 & \hat{\pi}_{it} s_{it} \\ \hat{\pi}_{it} s_{it} & s_{it}^2 \end{bmatrix}}_{K} \begin{bmatrix} \delta_i \\ \beta_i \end{bmatrix} \right)$$

Assuming the instrument is correlated with π_{it}, then it must be correlated with δ_i or β_i, as it cannot be correlated with s_{it}. Thus, K does not factor out. This means that again we do not recover $\mathbb{E}(\delta_i)$ or $\mathbb{E}(\beta_i)$.

C.2 Models estimated with simulated maximum likelihood

I estimate models with one signal weight, and two signals weights where conflicting and non-conflicting signals each have their own weight.

C.2.1 Models with one signal weight β_i and prior weight δ_i

1. Baseline model

$$\pi_{it} = \beta_i \lambda_{it} + \delta_i \pi_{i,t-1} + u_{it} \tag{6}$$

2. Mean shift β_t for each $t > 1$

$$\pi_{it} = (\beta_i + \beta_t) \lambda_{it} + \delta_i \pi_{i,t-1} + u_{it} \tag{7}$$

3. Time-specific variance for β_i for each $t > 1$

The time specific variance is defined as follows. For each time, we define a parameter κ_t that scales the covariance matrix of the parameter vector θ_i. I will drop the t to reduce notation. Let:

$$\theta_i = \begin{bmatrix} \beta_i \\ \delta_i \end{bmatrix}, \qquad \text{Cov}(\theta_i) = \begin{bmatrix} \sigma_{\beta_i} & \sigma_{\beta_i, \delta_i} \\ \sigma_{\beta_i, \delta_i} & \sigma_{\delta_i} \end{bmatrix}.$$

Let the scaling matrix be diagonal:

$$C = \begin{bmatrix} \sqrt{\kappa} & 0 \\ 0 & 1 \end{bmatrix}, \qquad y_i = C\theta_i = \begin{bmatrix} \sqrt{\kappa}\, \beta_i \\ \delta_i \end{bmatrix}.$$

Then

$$\begin{aligned}\mathrm{Cov}(y_i) &= C\,\mathrm{Cov}(\theta_i)\,C \\ &= \begin{bmatrix} \sqrt{\kappa} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{\beta_i} & \sigma_{\beta_i,\delta_i} \\ \sigma_{\beta_i,\delta_i} & \sigma_{\delta_i} \end{bmatrix} \begin{bmatrix} \sqrt{\kappa} & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \kappa\,\sigma_{\beta_i} & \sqrt{\kappa}\,\sigma_{\beta_i,\delta_i} \\ \sqrt{\kappa}\,\sigma_{\beta_i,\delta_i} & \sigma_{\delta_i} \end{bmatrix}.\end{aligned}$$

In other words, this scales the variance by κ and scales the covariance in column j and row k by $C_j C_k$ in column j and row k. The model estimated is:

$$\pi_{it} = \beta_i \lambda_{it} + \delta_i \pi_{i,t-1} + u_{it} \tag{8}$$

4. Mean shift β_t for each $t > 1$; time-specific variance for β_i for each $t > 1$

$$\pi_{it} = (\beta_i + \beta_t)\lambda_{it} + \delta_i \pi_{i,t-1} + u_{it} \tag{9}$$

C.2.2 Models with two signal weights β_{in} and β_{ic} and prior weight δ_i

Each model below is estimated with two definitions of a conflicting signal:

1. A conflicting signal is any signal where the correct update is in the opposite direction of the previous correct update (e.g., in $\{0, 1, 1, 0\}$, the first 1 and final 0 are conflicting).

2. A conflicting signal is the first signal in the sequence that conflicts (e.g., in $\{0, 1, 1, 0\}$, the first 1 is conflicting).

1. Baseline model with β_{in}, β_{ic}, and δ_i

$$\pi_{it} = \beta_{in}\lambda^n_{it} + \beta_{ic}\lambda^c_{it} + \delta_i \pi_{i,t-1} + u_{it} \tag{10}$$

2. Mean shift β_t for each $t > 1$

$$\pi_{it} = (\beta_{in} + \beta_t)\lambda^n_{it} + (\beta_{ic} + \beta_t)\lambda^c_{it} + \delta_i \pi_{i,t-1} + u_{it} \tag{11}$$

3. Mean shifts β_{tn} and β_{tc} for each $t > 1$

$$\pi_{it} = (\beta_{in} + \beta_{tn})\lambda_{it}^{n} + (\beta_{ic} + \beta_{tc})\lambda_{it}^{c} + \delta_i \pi_{i,t-1} + u_{it} \tag{12}$$

4. Time-specific variance for β_{in} and β_{ic} for each $t > 1$

The time specific variance is defined in the same way as the 1 weight case. For each time, we define a parameter κ_t that scales the covariance matrix of the parameter vector θ_i. I will drop the t to reduce notation. Let:

$$\theta_i = \begin{bmatrix} \beta_{in} \\ \beta_{ic} \\ \delta_i \end{bmatrix} \qquad \text{Cov}(\theta_i) = \begin{bmatrix} \sigma_{\beta_{in}} & \sigma_{\beta_{in},\beta_{ic}} & \sigma_{\beta_{in},\delta_i} \\ \sigma_{\beta_{in},\beta_{ic}} & \sigma_{\beta_{ic}} & \sigma_{\beta_{ic},\delta_i} \\ \sigma_{\beta_{in},\delta_i} & \sigma_{\beta_{ic},\delta_i} & \sigma_{\delta_i} \end{bmatrix},$$

$$C = \begin{bmatrix} \sqrt{\kappa} & 0 & 0 \\ 0 & \sqrt{\kappa} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad y_i = C\theta_i = \begin{bmatrix} \sqrt{\kappa}\,\beta_{in} \\ \sqrt{\kappa}\,\beta_{ic} \\ \delta_i \end{bmatrix}.$$

Then:

$$\begin{aligned} \text{Cov}(y_i) &= C\ \text{Cov}(\theta_i)\, C \\ &= \begin{bmatrix} \sqrt{\kappa} & 0 & 0 \\ 0 & \sqrt{\kappa} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{\beta_{in}} & \sigma_{\beta_{in},\beta_{ic}} & \sigma_{\beta_{in},\delta_i} \\ \sigma_{\beta_{in},\beta_{ic}} & \sigma_{\beta_{ic}} & \sigma_{\beta_{ic},\delta_i} \\ \sigma_{\beta_{in},\delta_i} & \sigma_{\beta_{ic},\delta_i} & \sigma_{\delta_i} \end{bmatrix} \begin{bmatrix} \sqrt{\kappa} & 0 & 0 \\ 0 & \sqrt{\kappa} & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \kappa\,\sigma_{\beta_{in}} & \kappa\,\sigma_{\beta_{in},\beta_{ic}} & \sqrt{\kappa}\,\sigma_{\beta_{in},\delta_i} \\ \kappa\,\sigma_{\beta_{in},\beta_{ic}} & \kappa\,\sigma_{\beta_{ic}} & \sqrt{\kappa}\,\sigma_{\beta_{ic},\delta_i} \\ \sqrt{\kappa}\,\sigma_{\beta_{in},\delta_i} & \sqrt{\kappa}\,\sigma_{\beta_{ic},\delta_i} & \sigma_{\delta_i} \end{bmatrix}. \end{aligned}$$

The model is estimated as:

$$\pi_{it} = \beta_{in}\lambda_{it}^{n} + \beta_{ic}\lambda_{it}^{c} + \delta_i \pi_{i,t-1} + u_{it} \tag{13}$$

5. Mean shift β_t for each $t > 1$; time-specific variance for β_{in} and β_{ic} for each $t > 1$

$$\pi_{it} = (\beta_{in} + \beta_t)\lambda_{it}^{n} + (\beta_{ic} + \beta_t)\lambda_{it}^{c} + \delta_i \pi_{i,t-1} + u_{it} \tag{14}$$

6. Mean shifts β_{tn} and β_{tc} for each $t > 1$; time-specific variance for β_{in} and β_{ic} for each $t > 1$

$$\pi_{it} = (\beta_{in} + \beta_{tn})\lambda_{it}^{n} + (\beta_{ic} + \beta_{tc})\lambda_{it}^{c} + \delta_i \pi_{i,t-1} + u_{it} \tag{15}$$

References

Agranov, M., & Reshidi, P. (2024). *Disentangling suboptimal updating: Task difficulty, structure, and sequencing*.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of behavioral economics: Applications and foundations 1* (pp. 69–186, Vol. 2).

Bland, J., & Rosokha, Y. (2025). *Rounding the (non)bayesian curve: Unraveling the effects of rounding errors in belief updating*.

Chan, K. (2025). *An axiomatic model and test of grether (1980) and bayes' rule* [SSRN working paper]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5084221

Danz, D., Vesterlund, L., & Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*.

Esponda, I., Vespa, E., & Yuksel, S. (2024). Mental models and learning: The case of base-rate neglect. *American Economic Review*, *114*(3).

Gonçalves, D., Libgober, J., & Willis, J. (2025). Retractions: Updating from complex information [Advance article]. *The Review of Economic Studies*. https://doi.org/10.1093/restud/rdaf032

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

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

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

Hsiao, C., & Pesaran, M. H. (2004). *Random coefficient panel data models*.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237.

Kieren, P., Müller-Dethard, J., & Weber, M. (2025). *Disconfirming information and overreaction in expectations*.

Lee, W. (2025). *Identification and estimation of dynamic random coefficient models*.

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

Raymond, C., & Wittrock, L. F. (2024). *Optimal memory with sequential learning: Signals or posterior beliefs* [SSRN working paper]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4765215