Online Appendix for: Disentangling Suboptimal Updating: Task Difficulty, Structure, and Sequencing

Marina Agranov Pëllumb Reshidi

August 12, 2026

Contents

1 Instructions

1.1 Common Instructions

The instructions shown in this subsection were seen by participants regardless of their treatment. These initial instructions aimed to familiarize participants with the mechanism through which they submitted their posteriors.

Figure 1: Initial Instructions I

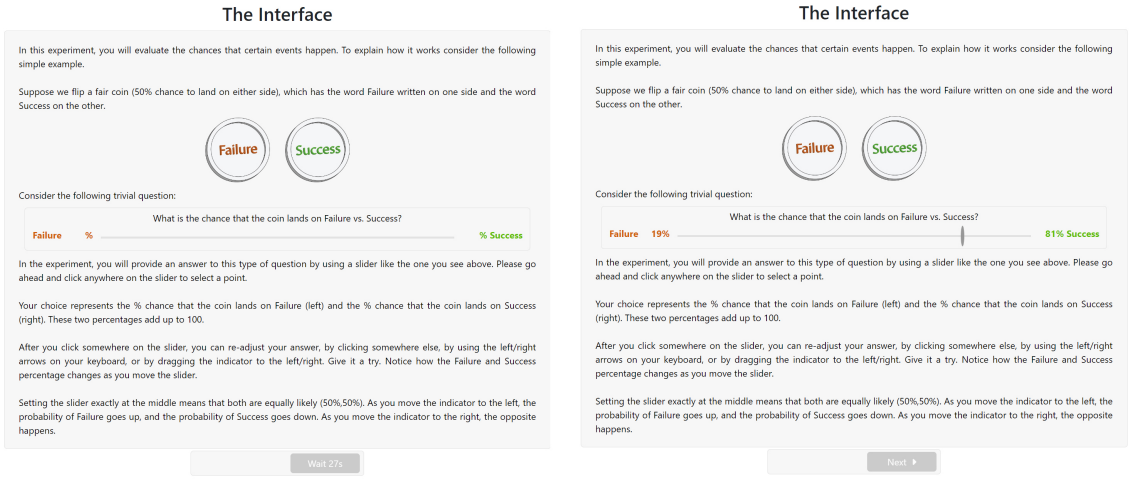
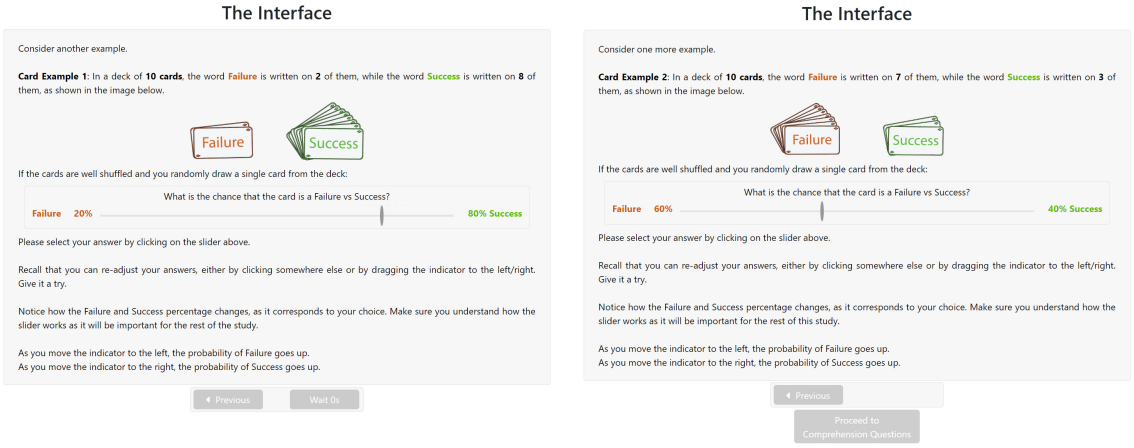


Figure 1 shows the initial page presented to the participants. To ensure that participants spend time internalizing the information, the *Next* button was made available only after a countdown of 30 seconds.[1] On this, and every other page, there is initially no indicator on the slider via which participants submit their probabilities. We made this decision to prevent participants from being anchored. The indicator and accompanying probabilities show up only after participants click somewhere on the slider. Compare the left (before clicking) and right (after clicking) screenshots in Figure 1.

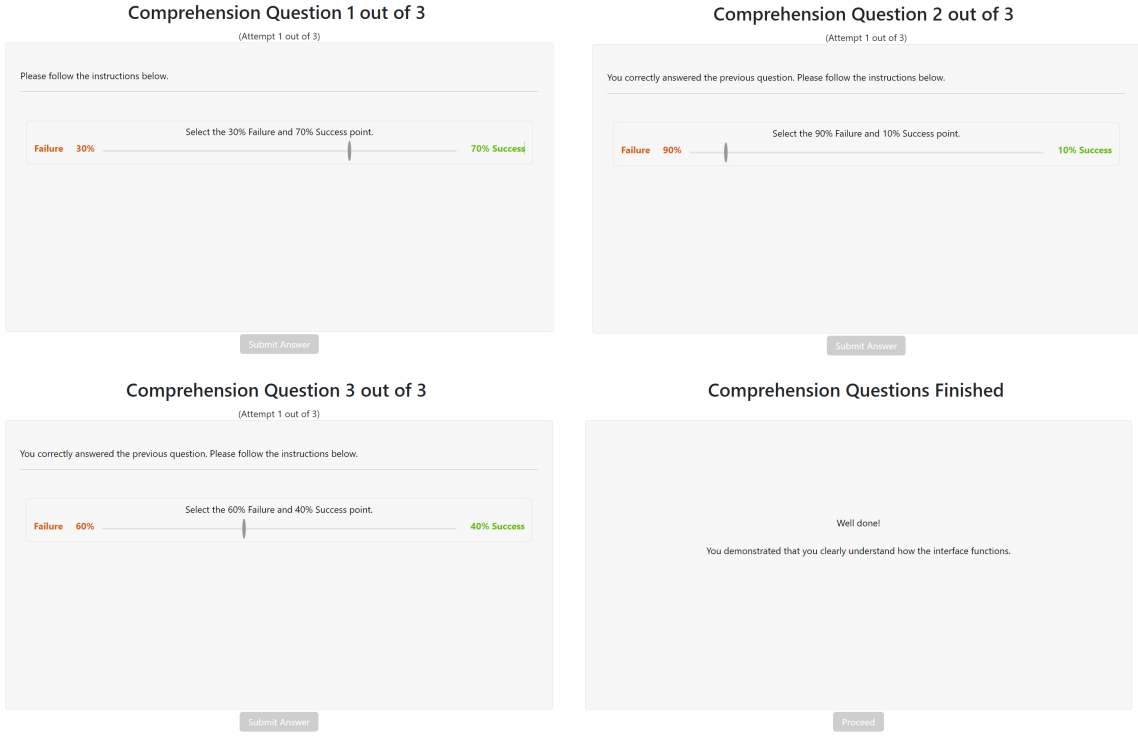
[1] Compare buttons on the bottom of the left and right screenshots shown in Figure 1. The left screenshot is taken 3 seconds after the page was loaded, whereas the right screenshot is taken after at least 30 seconds.

Figure 2: Initial Instructions II



Instructions continue by giving participants two more examples and reminding them how the mechanism works; see Figure 2. After these examples, participants are invited to start a simple comprehension test to ensure they know how to use the slider properly, see Figure 3.

Figure 3: Initial Instructions III

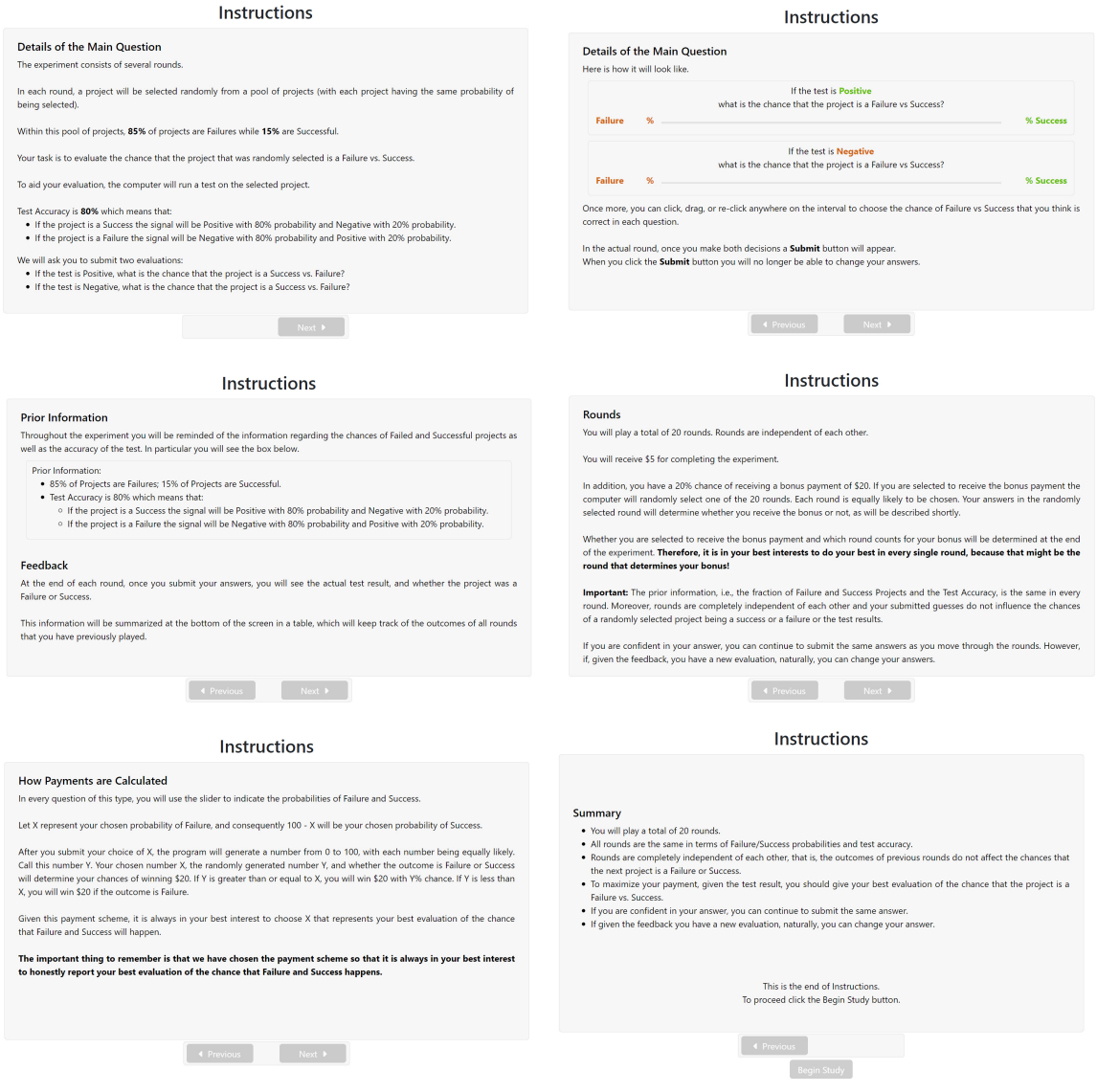


If participants submitted wrong answers more than twice, they were not al-

lowed to continue the study. Successful participants continued with treatment-specific instructions.

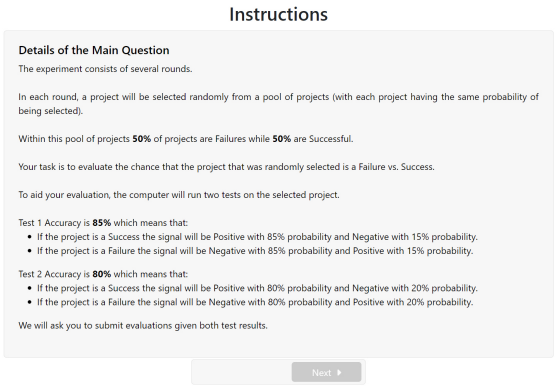
1.2 Baseline Treatment Instructions

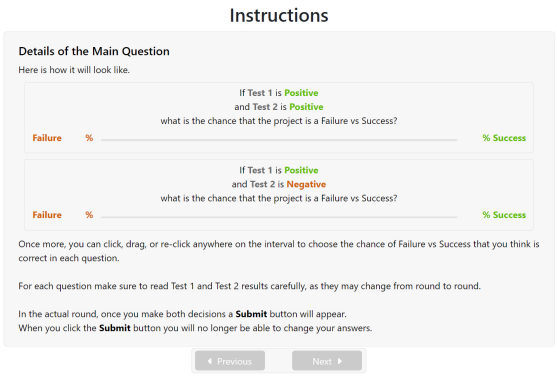
Figure 4: Baseline Treatment Instructions

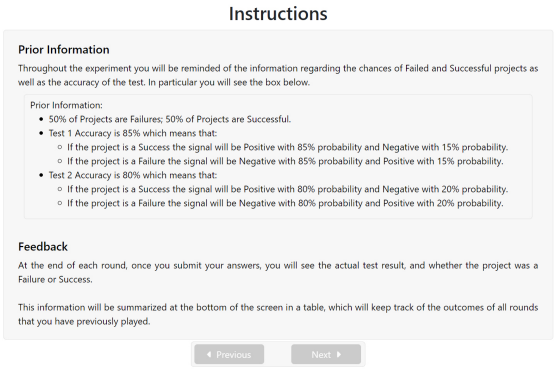


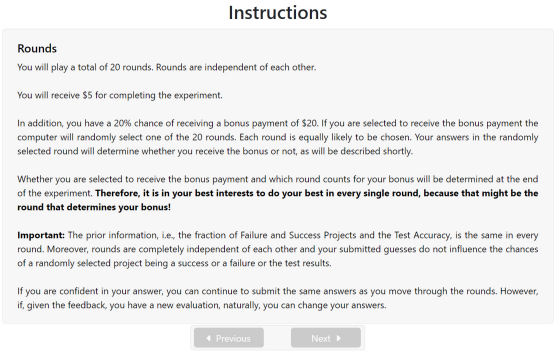
1.3 Simultaneous Treatment Instructions

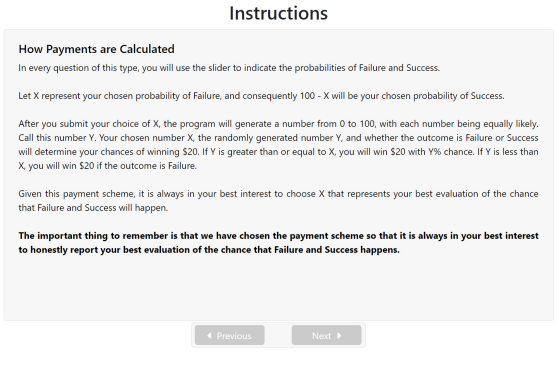
Figure 5: Simultaneous Treatment Instructions

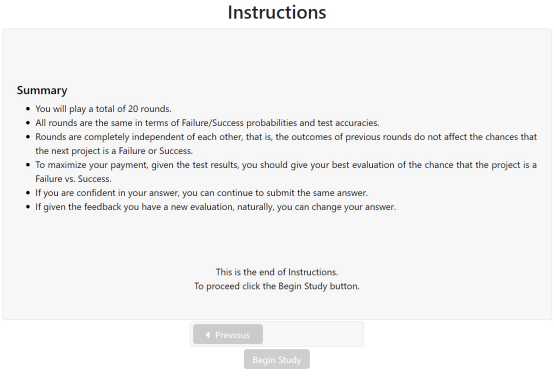






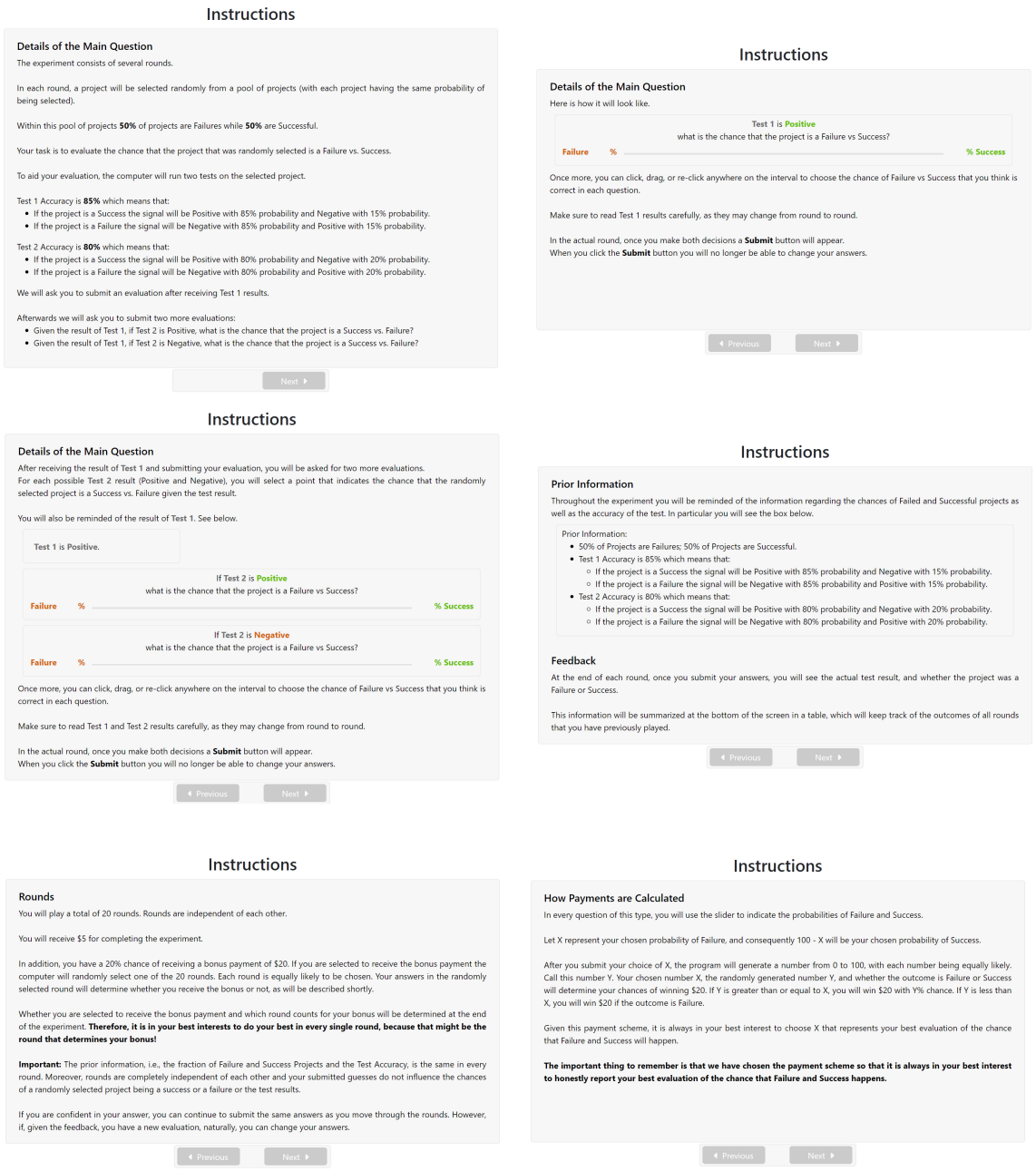






1.4 Sequential Treatment Instructions

Figure 6: Sequential Treatment Instructions

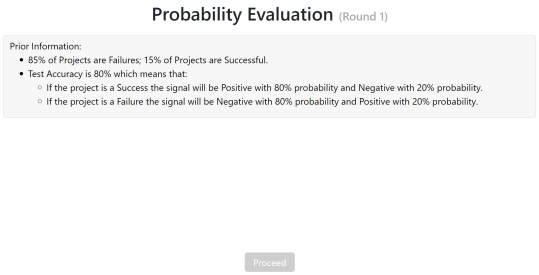


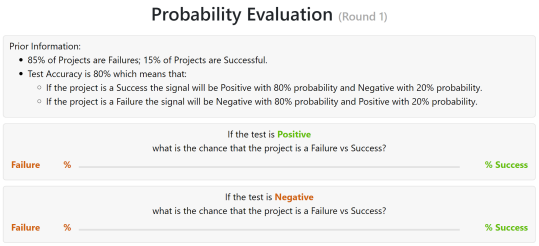
2 Interface

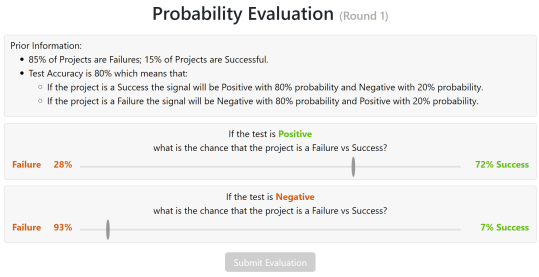
2.1 Baseline Treatment Interface

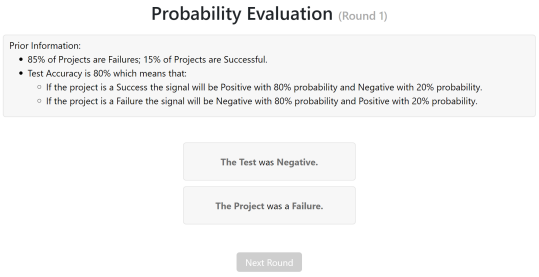
We present various screenshots of the interface presented to participants in the Baseline treatments at different stages of the study. We highlight important features below.

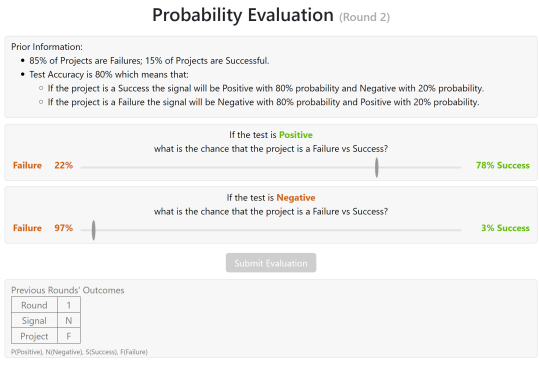
Figure 7: Baseline Treatment Interface

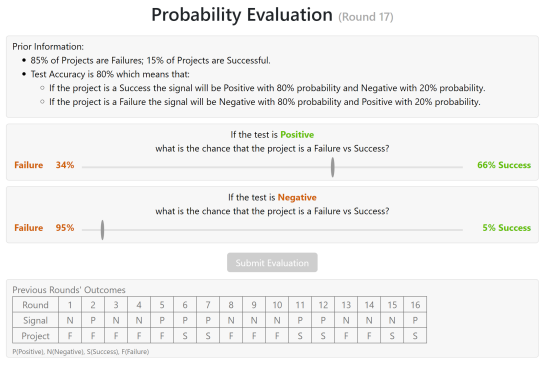






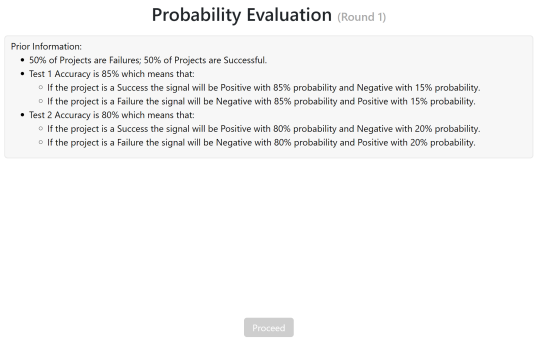


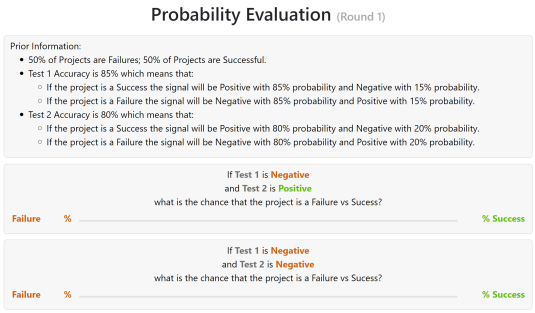


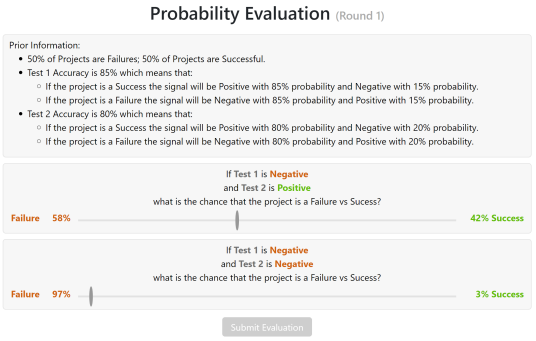


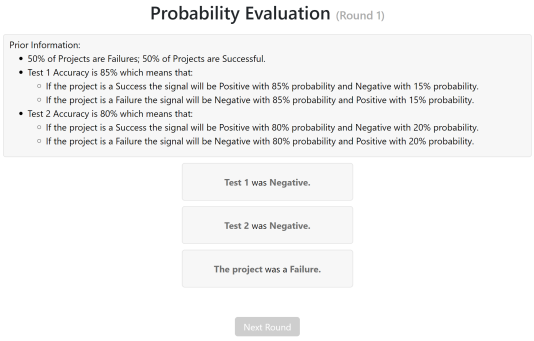
- As clarified in the instructions, throughout the experiment, at the top, participants see information regarding the prior probability of successful/failed projects as well as the signal accuracy.
- As clarified in the instructions, when asked *"If the test is Positive/Negative, what is the chance that the project is a Failure vs. Success?"* there is initially no indicator on the slider. We made this decision to prevent participants from being anchored. Only after they click somewhere on the slider does the indicator and the accompanying probabilities show up. For a concrete example, compare the top right and middle left screenshots in Figure 7.
- After clicking the "Submit Evaluation" button, participants were informed about the particular realized value of the signal and whether the project was a Failure or Success. See the middle right screenshot above.
- The realized signals and project outcomes from previous rounds are summarized in a table at the bottom of the interface. See bottom left for an example in Round 2 and bottom right for an example in Round 17. We keep track of past outcomes to shut down possible effects that imperfect recall may have.

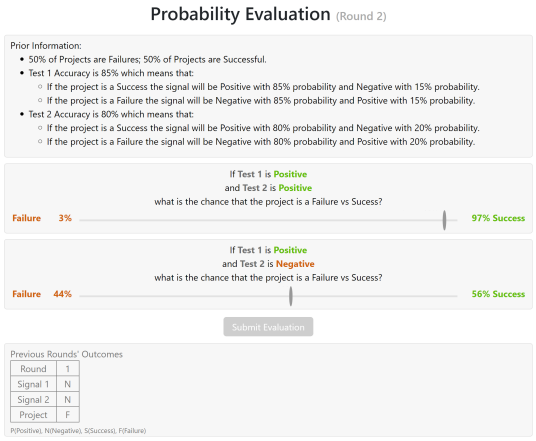
2.2 Simultaneous Treatment Interface

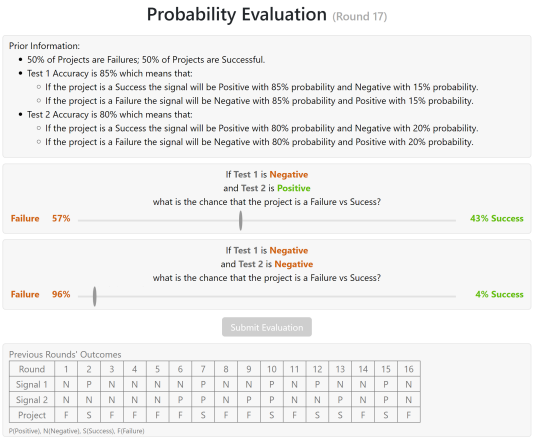






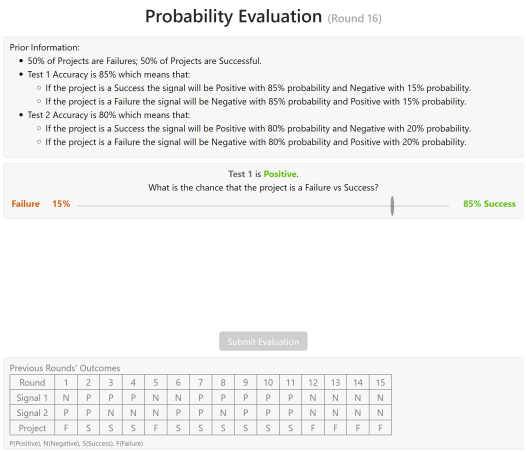


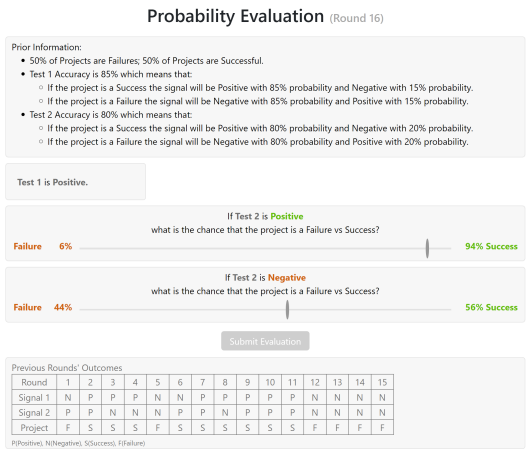




- Most of the design choices are unchanged from the Baseline treatment. However, in the simultaneous treatment, participants received both signals at the same time.

2.3 Sequential Treatment Interface





- Once more, most of the design choices are unchanged from the previous treatments. However, in the sequential treatment, participants received signals sequentially. Upon receiving the first signal, their posterior probability was elicited. Afterward, participants stated their posteriors conditional on the realized value of the second signal.

- The interface displays the outcome of the first signal when participants make choices conditional on the outcome of the second signal.

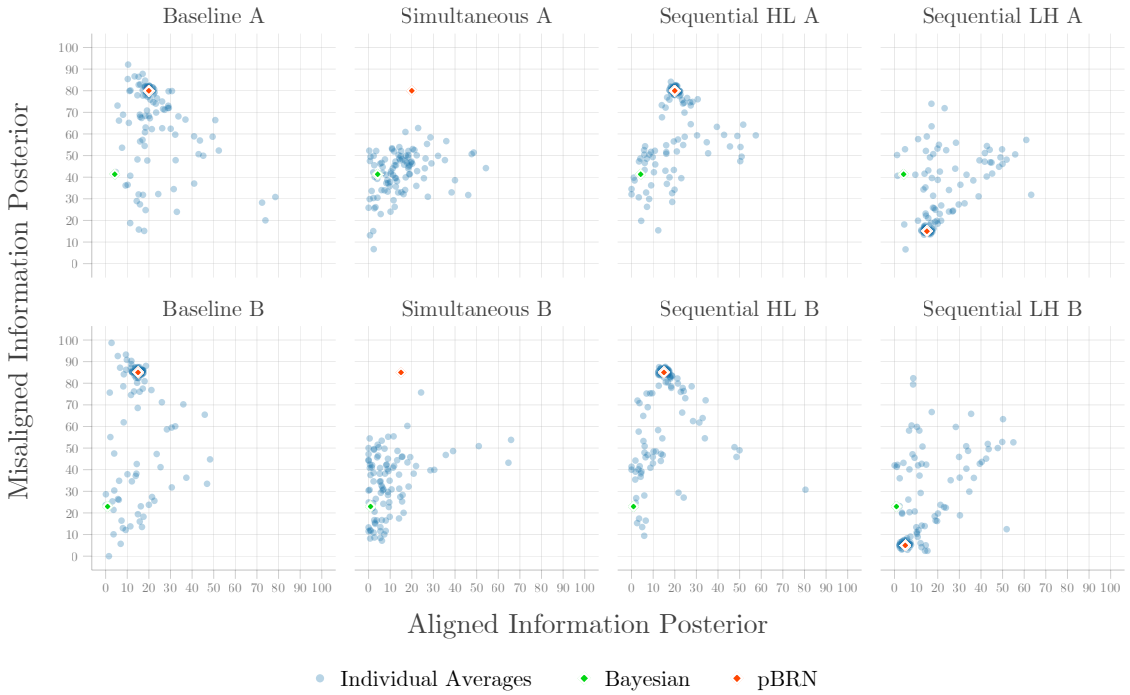
3 Individual Level Analysis

3.1 Primary Data Patterns

We now shift our attention to individual-level behavior. For each participant, we calculate their average elicited beliefs across all rounds for both aligned and misaligned information and present these averages in Figure 8.[2] Therefore, each datapoint in the figure represents the average behavior of a single participant.

[2] Figure 9 below shows the counterpart of Figure 8 using only the last five rounds. All main features remain unchanged.

Figure 8: Average Individual Choices



Notes: To help distinguish the large amount of data bundled on the pBRN level, we apply a jitter of 1.5 magnitude. This jittering perturbs the datapoint no further than a distance of 1.5 from the initial value. The top(bottom) row displays data across treatments under parametrization A(B).

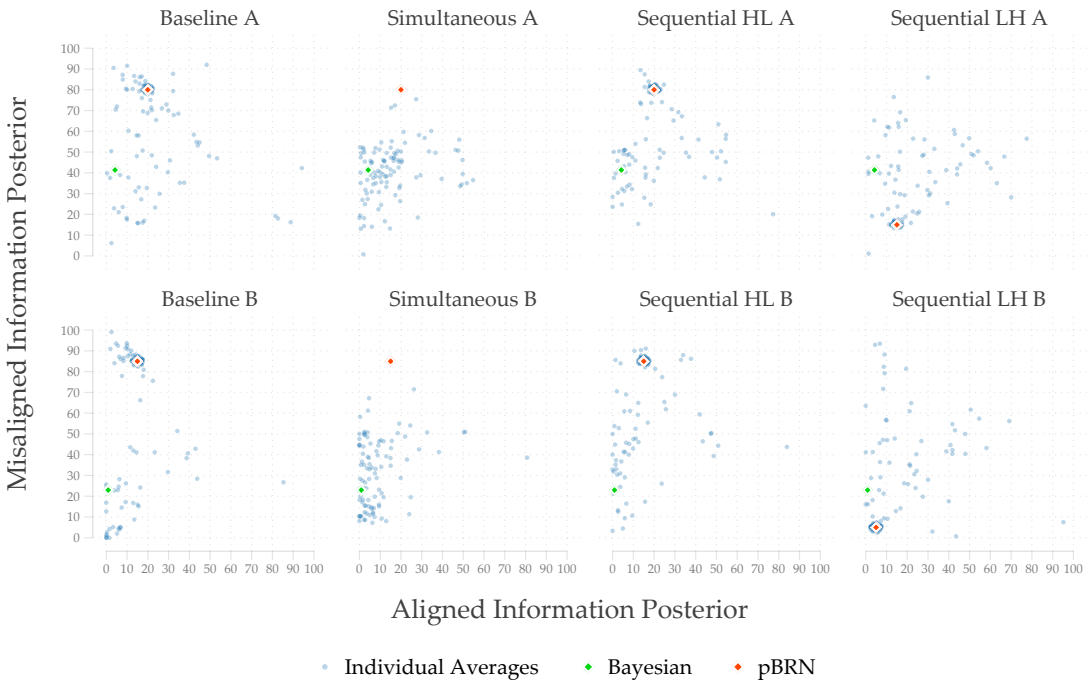
As can be seen, whenever information is released sequentially, the individual-level average posteriors are heavily bunched around the pBRN level.[3] This bunching phenomenon persists regardless of whether the sequential information delivery stems from an informative prior and a single signal (Baseline treatment) or an uninformative prior and two signals (Sequential treatments). Only when information is released simultaneously do we observe beliefs that are not heavily concentrated around the pBRN levels. These findings align with **Result 9** (Drivers of Base-Rate Neglect) in the main text, demonstrating the substantial impact of recency bias.

Result 1 (De-Bundling)**.** *The Simultaneous treatment is the only treatment leading to individual-level beliefs that are not strongly concentrated around the pBRN level.*

[3]Note that, given our normalization, in Sequential LH, the pBRN level is (15,15) under the first parametrization and (5,5) under the second.

Figure 9 displays the counterpart of Figure 8 using data from the last five rounds only. As can be seen, the takeaways are unchanged.

Figure 9: Average Individual Choices: Last Five Rounds



Notes: To help distinguish the large amount of data bundled on the pBRN level, we apply a jitter of 1.5 magnitude. This jittering perturbs the datapoint no further than a distance of 1.5 from the initial value. The top(bottom) row displays data across treatments under parametrization A(B).

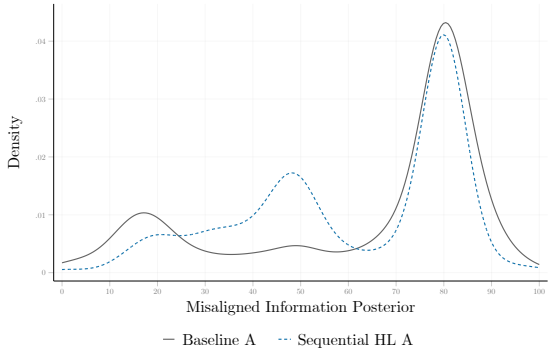
3.2 Individual-Level Effect of Information Structure

Recall that we observe no significant differences between average beliefs reported in the Baseline and the Sequential HL treatments in both parametrizations A and B (**4.4 Information Structure and Sequencing** in the main text). Although this holds true on average, in this section, we explore whether the information structure plays a role on an individual level.

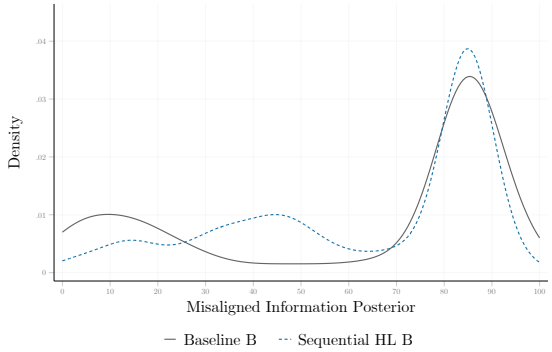
In Figure 10a and Figure 10b, we show estimated kernel densities of posteriors from the Baseline and Sequential HL treatments, under parametrization A and B respectively. Within a parametrization, the estimation pools posteriors across

Figure 10: Distribution of Posteriors and Individual-Level Averaged Posteriors

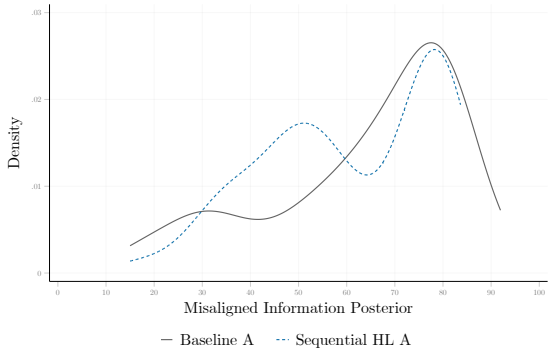
(a) Posteriors: Parameter A



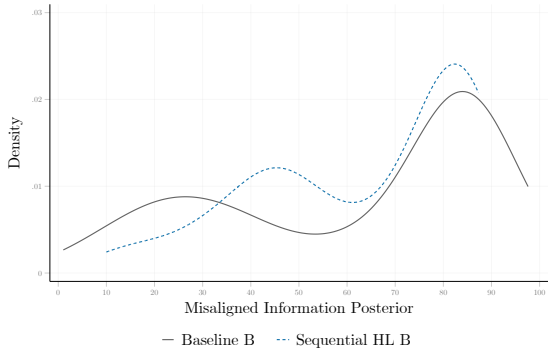
(b) Posteriors: Parameter B



(c) Individual Average Posteriors: Parameter A



(d) Individual Average Posteriors: Parameter B



participants and rounds.[4] In both parametrizations, despite the similar means, the distributions exhibit notable differences. In the Sequential HL treatments, there is a greater concentration toward intermediate values, which are close to the Bayesian level. While the fractions of participants choosing posteriors around the pBRN levels (80 for A and 85 for B) are comparable, in the Sequential HL treatment we see fewer values above these levels and fewer values for low posteriors. In both parametrizations, this mass is redistributed from the more extreme values toward the center in such a way that keeps the mean roughly unchanged. However, we run a Kolmogorov–Smirnov test between the Baseline and Sequential HL distributions and reject the null that they are the same ($p < 0.01$) under both parametrizations A and B.

The change between the distributions can be due to small changes in the behavior of many participants, drastic changes in the behavior of some participants,

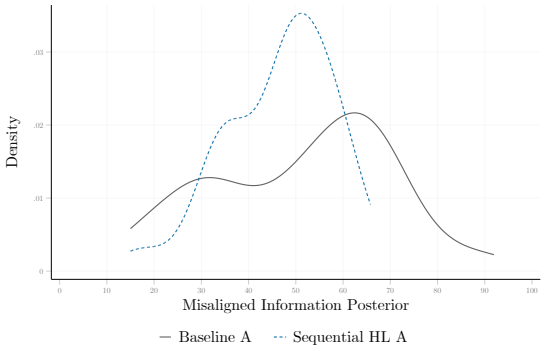
[4]Due to the consistent behavior exhibited by participants, conditioning on any round results in a qualitatively indistinguishable graph.

or both. To further explore this, we compute the average posterior for each participant across the 20 rounds and estimate kernel densities based on these average posteriors. Doing so allows us to focus on the variation across participants. In Figure 10c and Figure 10d, we present estimated kernel densities of the average individual-level posteriors in the Baseline and Sequential HL treatments, under parametrization A and B respectively. As can be seen, a considerable portion of participants, on average, choose levels near the pBRN levels (80 for A and 85 for B). These participants disregard the initial information and solely follow the second signal. For the remainder of the distributions, we see that in the Sequential HL treatment, fewer individuals choose extreme values. Our interpretation of these additional estimated kernel densities is that changing the information structure has no effect on individuals who solely follow the second signal, while for others, it steers their choices toward less extreme values—closer to the Bayesian level.

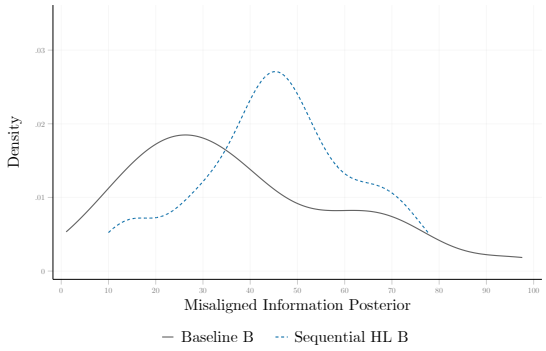
Below, we estimate the kernel densities under both parametrizations after removing participants that seem to behave in a pBRN manner. We remove participants from the analysis if their average aligned and misaligned posteriors are within 10 points from the pBRN posterior level.[5]

Figure 11: Individual-Level Averaged Posteriors Excluding pBRN

(a) Parametrization A



(b) Parametrization B



Compared to Figure 10a and Figure 10b, the difference between the estimated kernel densities becomes starker. This is in line with our interpretation that in-

[5]For parametrization A, this implies we drop participants whose misaligned posteriors are between 70 and 90 and whose aligned posteriors are between 10 and 30. For parametrization B, this implies we drop participants whose misaligned posteriors are between 75 and 95 and whose aligned posteriors are between 5 and 25.

formation structure seems to have minimal to no effect on participants who show pBRN behavior; however, for other non-pBRN participants, the effect is sizable. This is in line with our interpretation of the differential effect of information structure. In other words, for participants who only focus on the most recent information, the particular information structure does not play a substantial role—they ignore the initial information regardless. On the other hand, for participants who somewhat incorporate both the initial and the more recent information, the specific information structure can influence belief updating.

Result 2 (Effect of Information Structure)**.** *Elicited beliefs of participants who exclusively rely on recent information are unaffected by information structure. For other participants, a change in the information structure results in less extreme reported beliefs.*

3.3 Classifying Types: K-means Clustering

We next proceed by classifying participants into different types. To determine the number of types and the types themselves, we use k-means clustering, which, simply put, is a method that partitions n observations into k clusters/groups. Each observation is associated with the cluster with the nearest mean (centroid). This results in a partitioning of the data space into Voronoi cells. Specifically, k-means clustering minimizes within-cluster variances (squared Euclidean distances). This is one of the most commonly used unsupervised classifiers.[6] By employing this procedure, we bypass the need to determine types subjectively. Instead, we rely on the unsupervised classification procedure to determine both the number of types and their characteristics. To determine the number of clusters, we employ two commonly used approaches, the *elbow method* and the *silhouette score*. Details of these approaches are presented in the Section 3.4. Based on this initial analysis, the suggested number of clusters is three.

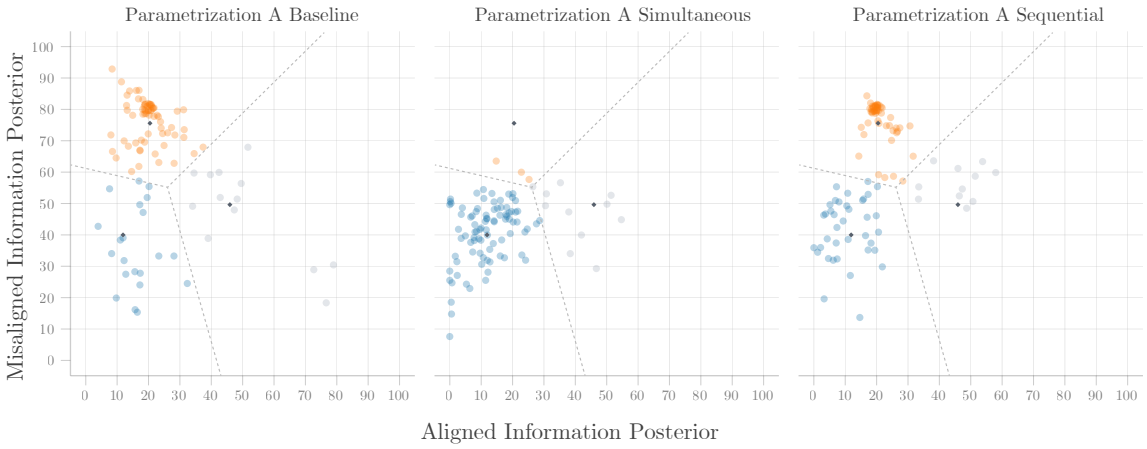
Since our aim is to evaluate how the share of different types changes across treatments, we separate the typical K-means clustering into two parts. The first part, clustering, involves determining the centroids for each cluster. We do this by pooling the data across treatments within a parametrization. Having determined

[6] An unsupervised classifier is a machine learning algorithm that automatically identifies patterns and groups data without prior labeled training examples.

the centroids, we then proceed with the second part, classification, which simply associates each observation to the cluster with the nearest mean.[7] To identify the centroids, we use a standard iterative refinement technique. To summarize, we do the following: (i) determine the number of clusters, (ii) determine the centroids, and (iii) classify participants.

We follow the exercise described above for treatments one through three. We exclude the Sequential LH treatment for technical reasons.[8] The clustered data is shown in Figure 12, along with the three centroids and the corresponding Voronoi sets they generate.

Figure 12: Parametrization A Clustering



Notes: Participants are categorized into three separate clusters. Dark gray dots mark the centroids of these clusters, and dashed lines represent the Voronoi cells corresponding to these centroids.

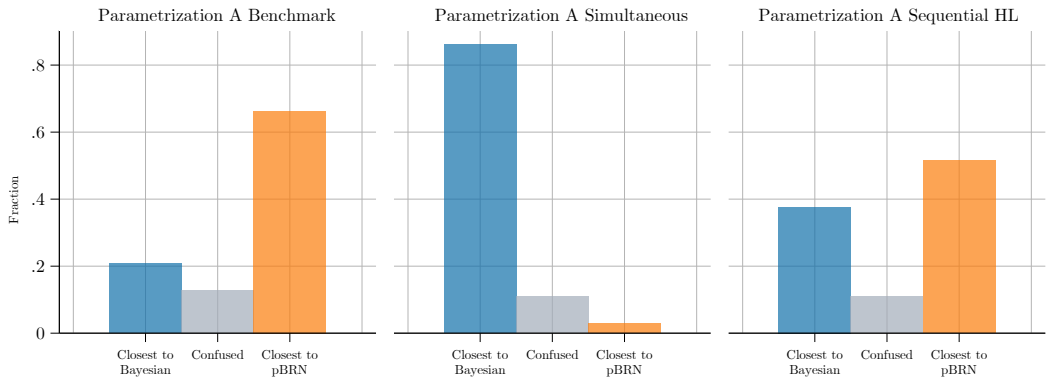
We see the emergence of three distinct clusters, with their centroids exhibiting close proximity to the Bayesian posterior (4.2,41.38), the 50-50 posterior, and the pBRN posterior (20,80). Consequently, we interpret the first group as roughly Bayesian, or closest to Bayesian, the second as potentially exhibiting confusion, and the third as roughly pBRN, or closest to pBRN. We label the second cluster as

[7] Had we not followed the procedure described above, and instead, had we estimated centroids for each treatment, there would be no natural way to compare shares of participants belonging to different groups across treatments since what a group is would differ from treatment to treatment.

[8] In the Sequential LH treatment, if we do not normalize the data, as we have done in the main analysis, the Bayesian posterior will have a different position compared to the three other treatments. If we normalize the data, the pBRN posterior will have a different position compared to the three other treatments. This, in turn, hampers our ability to have a natural interpretation of the clusters. Hence, we proceed with the clustering exercise for the first three treatments only.

potentially confused due to the fact that, regardless of the prior and signal value, whether positive or negative, participants consistently opt for values close to 50. We display the fraction of each type across treatments in Figure 13.

Figure 13: Parametrization A Cluster Histogram

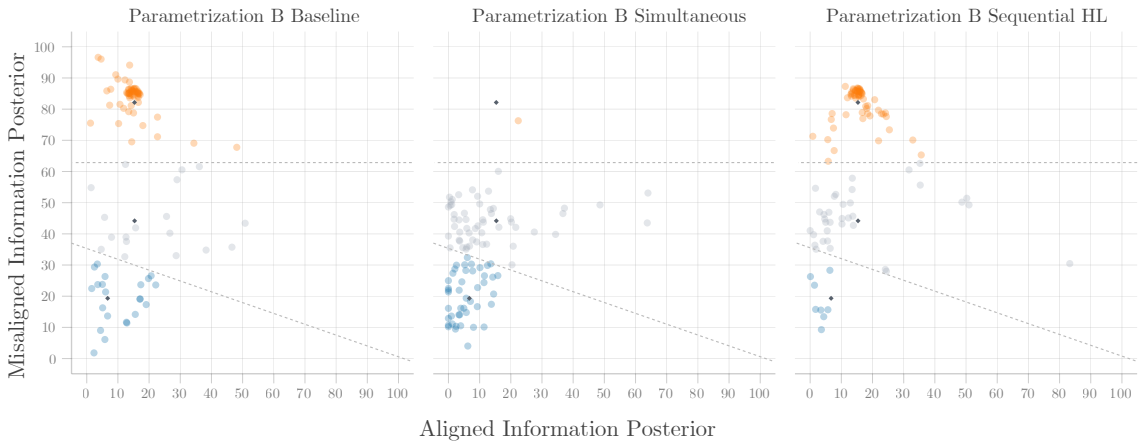


As can be seen, a simultaneous release of information under the first parametrization leads to the largest share of participants classified as closest to Bayesian, with a minuscule share of agents closest to the pBRN level. Importantly, although in the previous section, we saw that the estimated mean under the Baseline and Sequential HL treatments was not statistically different, we see that the composition of the type of participants differs. The Sequential HL treatment is characterized by a higher share of closest-to-Bayesian agents and a lower share of closest-to-pBRN agents. Thus, in line with the evidence presented in Section 3.2, the information structure does seem to have an effect on individual-level behavior.

Result 3 (Information Structure and Type Classification)**.** *Information structure affects participant categorization.*

We next turn our attention to parametrization B. We once again follow the procedure described above and show the clustered data in Figure 14, along with the three centroids and the corresponding Voronoi sets they generate.

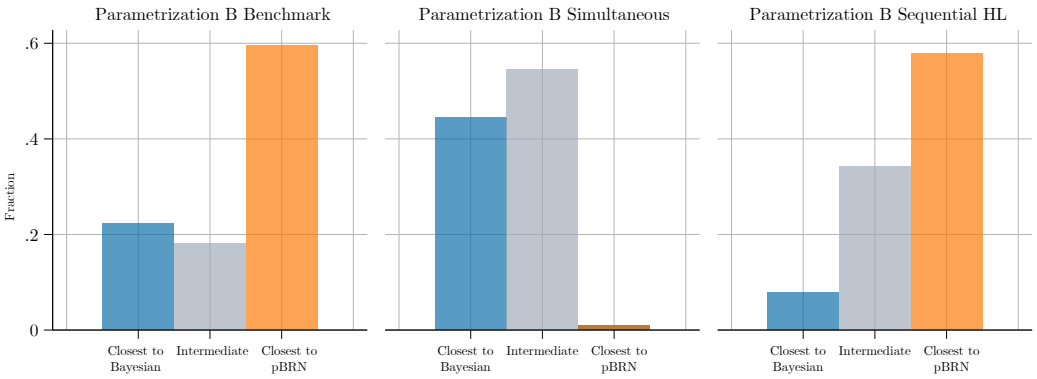
Figure 14: Parametrization B Clustering



Notes: Participants are categorized into three separate clusters. Dark gray dots mark the centroids of these clusters, and dashed lines represent the Voronoi cells corresponding to these centroids.

We see the emergence of three distinct clusters, with their centroids exhibiting close proximity to the Bayesian posterior (0.9,22.97), an in-between posterior, and the pBRN posterior (15,85). Consequently, we interpret the first group as roughly Bayesian, or closest to Bayesian, the second as in between the two extremes, and the third as roughly pBRN, or closest to pBRN. We display the fraction of each type across treatments in Figure 15.

Figure 15: Parametrization B Cluster Histogram



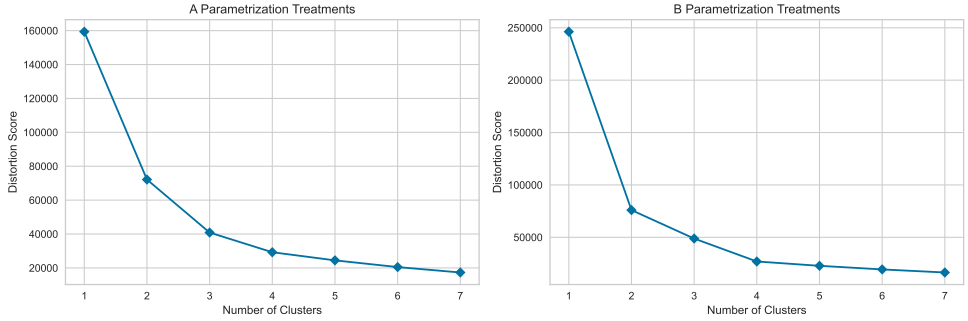
Once more, a simultaneous release of information leads to the largest share of participants classified as closest to Bayesian, with a minuscule share of agents classified as closest to the pBRN level. We once again see that the classification of participants differs between the Baseline and Sequential HL treatments.

Result 4 (Types Across Parameters)**.** *We observe variation in both clusters as well as the distribution of participants among these clusters across different parameters.*

3.4 Classifying Types

The elbow method is a way to determine the optimal number of clusters in a dataset for k-means clustering. It works by plotting the sum of squared distances between each point and the centroid of its cluster against the number of clusters used. The plot looks like an arm, and the elbow point on the arm represents the best number of clusters to use. This is because the elbow point is where adding more clusters does not significantly improve the clustering results. The elbow method helps to select an appropriate number of clusters for k-means clustering, avoiding underfitting or overfitting the data. The graphs shown in Figure 16 reveal that the elbow method recommends three clusters for parametrization A, while for parametrization B, the score is somewhat ambiguous between two, three, and four clusters. We supplement our calculations by determining the optimal number of clusters via the silhouette method.

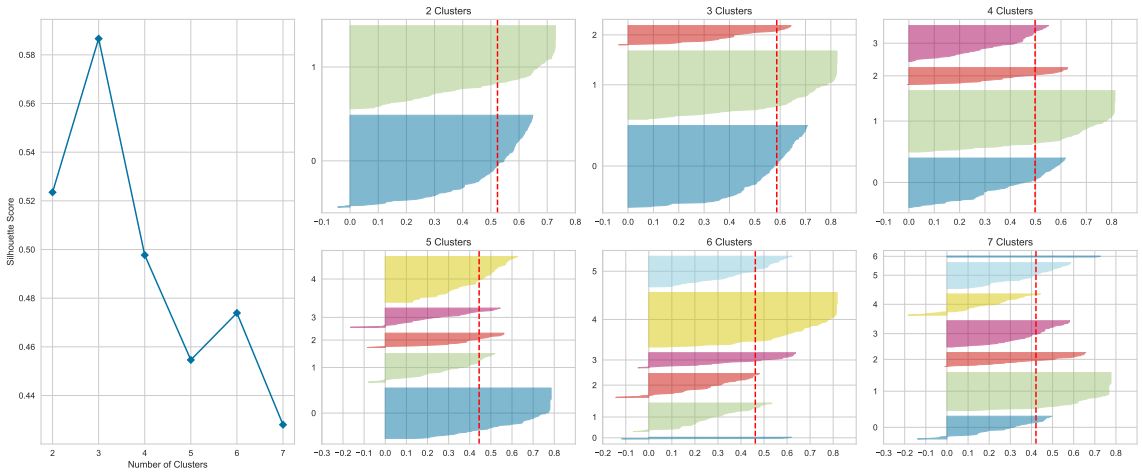
Figure 16: Distortion Score Elbow for K-Means Clustering



The silhouette method is a way to evaluate the quality of clustering results in a dataset. It works by measuring how similar an observation is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, with higher values indicating better clustering results. A score of 1 indicates that the observation is well-matched to its own cluster and poorly-matched to other clusters. A score of -1 indicates the opposite, while a score of 0 indicates that the observation is equally similar to its own cluster and other clusters. The silhouette method calculates the average silhouette score of all observations in the dataset and uses this

as a measure of how well the data is clustered. The method can be used to compare different clustering methods or to select the best number of clusters to use in a k-means clustering analysis. By selecting the number of clusters that maximizes the silhouette score, the method can help improve the accuracy and reliability of the clustering results. The graphs shown in Figure 17 reveal that the silhouette score is maximized under three clusters.

Figure 17: Silhouette Scores For K-Means Clustering



We therefore decide to proceed with the clustering exercise with three clusters.

4 Additional Analysis

4.1 Heterogeneity in Belief Updating Across Demographics

Recall from **Section 4.2** that we estimate the parameter α, which quantifies how well agents incorporate nonlinearities into their belief updating. A value of $\alpha = 0$ indicates a complete failure to account for nonlinearities, while $\alpha = 1$ corresponds to behavior indistinguishable from Bayesian updating. In this section, we repeat the same estimation exercise but examine whether the estimated values of α differ by sex, age, or ethnicity. As shown in **Table 3** of the main text, on average participants place approximately 42% weight on the Bayesian posterior, which rises to 56% in the final five rounds. This pattern suggests that participants move toward more Bayesian updating as they gain experience, though even late in the study

they remain far from fully incorporating the nonlinearities implied by Bayesian updating.

Table 1 below reveals that female participants are somewhat more Bayesian than males (approximately 0.50 versus 0.35), but the gap is not statistically significant in either specification. This pattern persists when we restrict attention to the final five rounds, where the estimates are approximately 0.61 for females and 0.52 for males. Thus, while the difference is notable in magnitude, sex does not emerge as a statistically significant predictor of updating behavior.

Table 1: Estimated α by Sex

	All Rounds	Last 5 Rounds
$\hat{\alpha}$ (Female)	0.495***	0.612***
	(0.0788)	(0.0964)
$\Delta\hat{\alpha}$ (Male)	-0.142	-0.0966
	(0.111)	(0.134)
N (observations)	11,780	2945
K (individuals)	589	589

Individual-level clustered errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Next, we split participants into terciles at ages 27 and 39, corresponding to the 33rd and 67th percentiles. As shown in Table 2, the youngest group (under 27) is the most Bayesian, with an estimated α of 0.64. The gap widens with age: the middle tercile (ages 27–38) scores 0.22 lower ($p < 0.10$), while the oldest tercile (ages 39 and above) scores 0.43 lower ($p < 0.01$). When we restrict attention to the final five rounds, all three groups improve and the middle tercile's gap becomes insignificant; however, the oldest tercile remains significantly less Bayesian (-0.37, $p < 0.05$). In short, older participants learn over the course of the experiment but do not catch up to their younger counterparts.

Table 2: Estimated α by Age Tercile

	All Rounds	Last 5 Rounds
$\hat{\alpha}$ (Youngest)	0.635***	0.715***
	(0.0934)	(0.110)
$\Delta\hat{\alpha}$ (Middle)	-0.221*	-0.104
	(0.126)	(0.151)
$\Delta\hat{\alpha}$ (Oldest)	-0.432***	-0.370**
	(0.141)	(0.171)
N (observations)	11,680	2920
K (individuals)	584	584

Individual-level clustered errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Finally, we examine estimated α by ethnicity. As shown in Table 3, White participants have a baseline α of 0.43 ($p < 0.01$). None of the interaction terms are statistically significant in either specification, despite some large point estimates (e.g., Mixed ethnicity at +0.29). In the final five rounds, the baseline α rises to 0.58 and some pattern of interaction terms shifts, but no coefficient reaches statistical significance. Small subgroup sizes severely limit statistical power in this analysis.

Table 3: Estimated α by Ethnicity

	All Rounds	Last 5 Rounds
$\hat{\alpha}$ (White)	0.428***	0.579***
	(0.0658)	(0.0772)
$\Delta\hat{\alpha}$ (Black)	-0.249	-0.220
	(0.173)	(0.230)
$\Delta\hat{\alpha}$ (Asian)	0.0393	-0.247
	(0.185)	(0.267)
$\Delta\hat{\alpha}$ (Mixed)	0.293	0.362
	(0.251)	(0.263)
$\Delta\hat{\alpha}$ (Other)	0.191	0.168
	(0.337)	(0.351)
N (observations)	11,720	2930
K (individuals)	586	586

Individual-level clustered errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

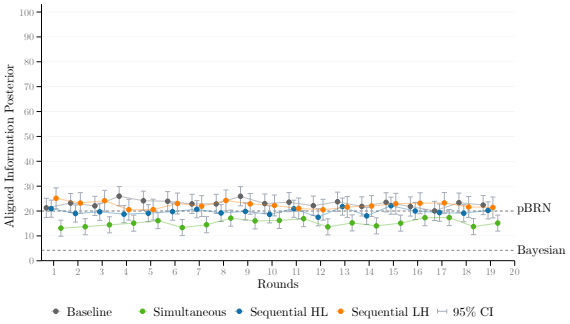
Overall, in our data, demographics appear to play a limited role in explaining variation in Bayesian reasoning. Only age shows a relatively robust and persistent association—older participants tend to be less Bayesian even after gaining experience with the task. Sex and ethnicity, by contrast, do not produce statistically significant differences in our sample, suggesting that the gap between Bayesian and linear updating may be largely uniform across demographic groups, though

we note that limited sample sizes in some subgroups constrain our ability to detect smaller effects.

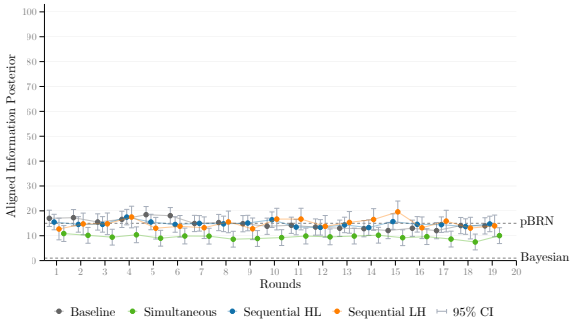
4.2 Aligned Information Posteriors

Figure 18: Aligned Information Posteriors

(a) Parametrization *A*



(b) Parametrization *B*



In Figure 18a and Figure 18b, we graph the average round-by-round beliefs of participants when information is aligned. For the Baseline treatment, these are cases when the realized signal is in the direction in which the prior leans. For all other treatments, these are cases in which both signals have the same realized value.

4.3 Pilot Data

Estimated Means We conducted two pilot studies under parametrization A for the Baseline and Simultaneous treatment. In Table 4, we compare the estimated mean from Baseline A and Simultaneous A with the estimated means in their corresponding pilot treatments. The variable *Constant* captures the estimated mean in the regular session, whereas the variable *Pilot* captures the difference of the estimated mean from this value in the pilot treatment. As can be seen, regardless of the error clustering level, the difference is never statistically significant.

Table 4: Estimated Means

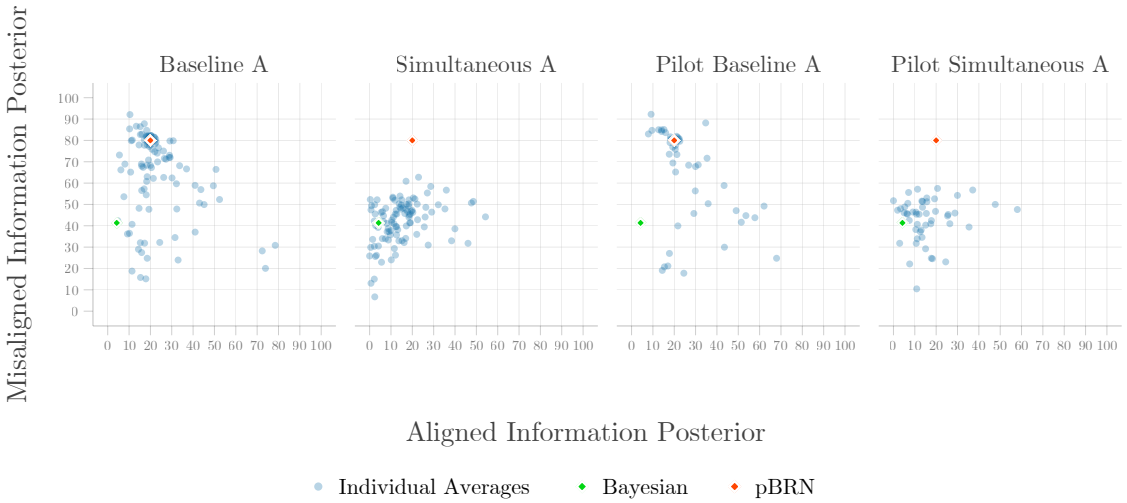
	Baseline A			Simultaneous A		
	No C	Ind C	Ind C + Last 5	No C	Ind C	Ind C + Last 5
Constant	63.79***	63.79***	60.43***	41.65***	41.65***	40.29***
	(0.595)	(1.971)	(2.428)	(0.384)	(0.987)	(1.295)
Pilot	0.434	0.434	5.911	1.001	1.001	1.483
	(1.041)	(3.680)	(4.285)	(0.667)	(1.740)	(2.382)
N	3,000	3,000	750	3,020	3,020	755

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Individual Level Analysis In Figure 19, we plot the individual level data for Baseline A and Simultaneous A, as well as their corresponding pilot treatments.

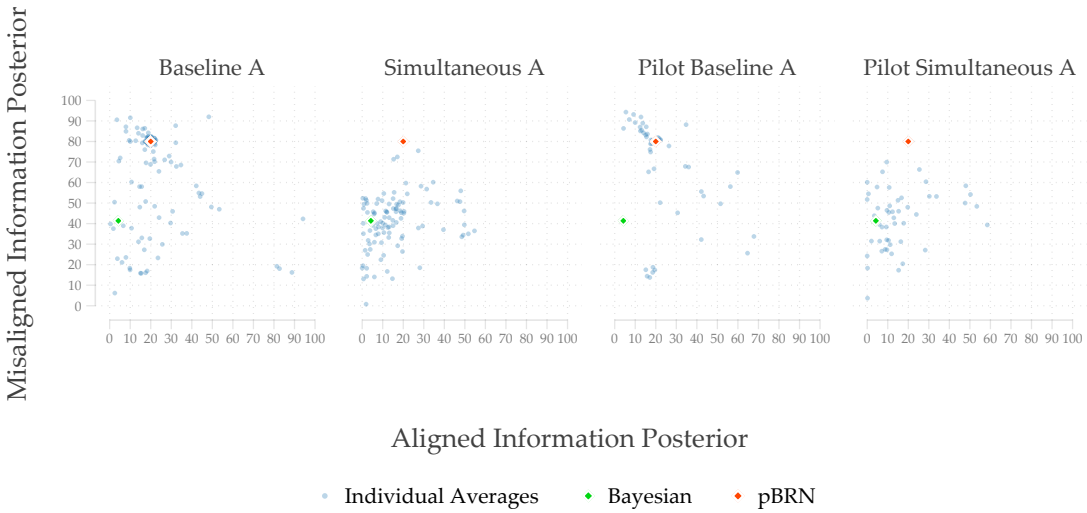
Figure 19: Average Individual Choices



Notes: To help distinguish the large amount of data bundled on the pBRN level, we apply a jitter of 1.5 magnitude. This jittering perturbs the datapoint no further than a distance of 1.5 from the initial value.

In Figure 20, we do the same using data from the last five rounds only.

Figure 20: Average Individual Choices: Last Five Rounds



Notes: To help distinguish the large amount of data bundled on the pBRN level, we apply a jitter of 1.5 magnitude. This jittering perturbs the datapoint no further than a distance of 1.5 from the initial value.