Supplementary Materials for

Learning *too* much from *too* little:
False face stereotypes emerge from a few exemplars and persist via insufficient sampling

Note. Please use this document in parallel with the codebook published on OSF (link) and the main text. This file is designed to be used as a reference for readers seeking information on specific topics.

Table of Contents:

S1. Experimental stimuli.

Face stimuli. We generated face stimuli using FaceGen3.1, a multidimensional computer-generated face space. We generated 200 identities, manipulated on shape dimensions from -10 SD to +10 SD, in increments of 2 SD. The manipulated dimension is orthogonal to any of the following core social dimensions: attractiveness, competence, dominance, extroversion, likeability, threat, and trustworthiness. The dimension, to a raw eye, turns out to reflect the width of a face. All face stimuli are uploaded in the OSF folder.

Behavior stimuli. In Experiment 1, we generated behavior stimuli using a trust game. In the trust game, we paired our participants with hypothetical partners indicated by face stimuli. Our participants were given 10 Monetary Units (MUs). They were told that they could send some of these MUs to an anonymous partner. They can send as few as 0 to as much as 10. They were also informed that whatever they send would be quadrupled; and then their partner would make a similar choice – send some of the MUs back to the participants. The upper bound of total earning, so the MUs a partner can send back, at each round was 20 and the lower bound was 0. The trust game is often used to measure cooperative motives in behavioral economics (e.g., participants who share more MUs are perceived as generous), but its structure is fairly complicated. To simplify the study for online workers, in Experiments 2, 3, 4, 5, and 6, we changed the trust game to a simple hypothetical event and asked participants to estimate generosity directly. The behavioral stimuli changed from $0 to $100, with an increment of $25 in Experiments 2 and 3, and with an increment of $50 in Experiments 4, 5, and 6.

Face-Behavior correlation. The correlation between face stimuli and behavior stimuli was predefined. The face variation ranged from -10 SD to +10 SD, in increments of 2 SD. The behavior variation ranged from 0 MU to 10 MU, in increments of 1 MU. In the population data (N=121 for Experiment 1 and N=25 for the remaining experiments), the linear relationship between face and behavior was close to zero ($r \approx 0.0$). In the initial sample (N=11 for Experiment 1, N=5 for Experiments 2 and 3, and N = 3 for Experiments 4, 5, and 6), we created a linearly positive, negative, or zero correlation.

For example, for the positive linear relationship, stimuli were arranged as the sequence of tuples: {-10SD, 0MU}, {-8SD, 1MU},{-6SD, 2MU},{-4SD, 3MU},{-2SD, 4MU},{0SD, 5MU},{2SD, 6MU},{4SD, 7MU},{6SD, 8MU},{8SD, 9MU}, {10SD, 10MU}. $r = + 1.0$. For the negative linear relationship, stimuli were arranged as the sequence of tuples: {10SD, 0MU}, {8SD, 1MU},{6SD, 2MU},{4SD, 3MU},{2SD, 4MU},{0SD, 5MU},{-2SD, 6MU},{-4SD, 7MU},{-6SD, 8MU},{-8SD, 9MU}, {-10SD, 10MU}. $r = - 1.0$. For the near zero linear relationship, stimuli were arranged as the sequence of tuples: {10SD, 4MU}, {8SD, 3MU},{6SD, 7MU},{4SD, 8MU},{2SD, 0MU},{0SD, 5MU},{-2SD, 10MU},{-4SD, 2MU},{-6SD, 6MU},{-8SD, 4MU}, {10SD, 9MU}. $r = -.09$.

For the 5 example stimuli, the positive condition was {-8SD, $0}, {-4SD, $25}, {0SD, $50}, {4SD, $75}, {8SD, $100}, $r = +1.0$; the negative condition was {-8SD, $100}, {-4SD, $75}, {0SD, $50}, {4SD, $25}, {8SD, $0}, $r = -1.0$; the near zero condition was {-8SD, $25}, {-4SD, $100}, {0SD, $0}, {4SD, $75}, {8SD, $50}, $r = + 0.1$.

For the 3 example stimuli, the positive condition was {-8SD, $0}, {0SD, $50}, {8SD, $100}, $r = +1.0$; the negative condition was {-8SD, $100}, {0SD, $50}, {8SD, $0}, $r = -1.0$; the near zero condition was {-8SD, $50}, {0SD, $0}, {4SD, $100}, $r = + 0.33$.

Regardless of the initial sample statistics, the population dataset's Pearson's correlation between face stimuli and behavior stimuli was always strictly $r = 0.00$.

S2. Experiment architecture.

Please read a detailed description of the procedure in the main text in Experiment 1 - procedures. Readers can access a version of the experiment with anonymized instructions [here]. For this live demonstration, we use materials from Experiment 6 where participants first see three initial examples, and then could freely decide whether to continue sampling or to jump to decisions. This is identical to Experiments 3 and 4. During the test phase, participants are provided with an opportunity to ask for additional information, which is a unique design feature of Experiment 6. If participants do not request more information, they proceed as others in Experiments 3 and 4. To reiterate from the main text, Experiment 5 varies the strength and the identity of the initial sampled faces and behaviors; Experiment 2 has five examples for the initial sample and the free versus fixed-sampling conditions. For a static presentation, we attach screenshots for convenience using Experiment 3 as an example [here]. Imagine this participant freely sampled 2 examples and went to the decision phase, which is illustrated with one test trial. Questionnaire and demographic questions are abbreviated in the screenshots and accessible via the above demonstration link.

S3. Analytical models.

The full model is a linear regression estimation with errors clustered at each participant, i.e., multi-level linear regression. To simplify the multilevel notations, the main model is denoted as the following:

$$y = X\beta + \varepsilon$$

where,

$y = \left(y_1 \; \dots \; y_n\right)$, dependent variable, trial-by-trial decisions from participant *1* to participant *n*. If 30 test faces, then each y_i has 30 values.

$X = \left(X_1^T \; \dots \; X_n^T\right) = \left(1 \; \dots \; 1 \; x_{11} \; x_{12} \; x_{13} \; \dots \; \dots \; \dots \; x_{n1} \; x_{n2} \; x_{n3} \; x_{14} \; x_{15} \; \dots \; \dots \; x_{n4} \; x_{n5}\right)$, predictor variables, 1 represents intercept; the first dimension, x_{i1}, represents manipulated facial dimension, continuous variable; the second dimension, x_{i2}, represents the initial sample condition, positive, negative, and zero, factor coded; the third dimension, x_{i3}, represents the sampling manipulation, free and fixed, factor coded; and the interaction between facial dimensions and initial sample condition, $x_{i4} = x_{i1} * x_{i2}$, as well as the interaction between facial dimensions and the sampling manipulation, $x_{i5} = x_{i1} * x_{i3}$.

$\beta = \left(\beta_0 \; \beta_1 \; \beta_2 \; \beta_3 \; \beta_4 \; \beta_5\right)$, linear estimators for the intercept (β_0), face dimensions (β_1), initial sample condition (β_2), sampling manipulations (β_3), face * initial interaction (β_4), and face * sampling interaction (β_5).

$\varepsilon = \left(\varepsilon_1 \; \dots \; \varepsilon_n\right)$, error term, clustered at each individual participant level, to account for participant-by-trial hierarchical structures.

In sum, the model tries to estimate the relationship between participants' responses (i.e., estimated donation) and three key experimental manipulations (i.e., facial features, initial sample, sampling, and interactions). β is estimated as the magnitude of the relationships;

confidence intervals, and p-values are used to compare, through repeated samples, how likely is that β is different from zero. Key results are the two interaction terms: participants' responses change as a function of different initial samples and as a function of different sampling processes.

For additional data processing, analysis details, as well as visualizations, please access Jupyter Notebook knitted in HTML for convenience viewing on OSF. Please also feel free to run the analysis. The analysis plan was pre-registered for all studies prior to data collection; below is a list of view-only pre-registration links:

Experiment 1: https://osf.io/qznug/?view_only=c8686a30c77f4dd0a31facbe43b5c631

Experiment 2: https://osf.io/gxuj5/?view_only=813d4151929b4fb68963fdc7daf59483

Experiment 4: https://osf.io/cwu7a/?view_only=a430a81db431490cba8457aebc6b35f6

Experiment 5: https://osf.io/ezgfq/?view_only=4e4c8aa8c8a948f8a63f7d2e01246ee7

Experiment 6: https://osf.io/7cjkg/?view_only=0572eacf9830406891c07f316d35fcd3

Note that Experiment 3 is the same as Experiment 2-free sampling condition, so no double registration.

S4. Supplementary results.

Individual heterogeneity. The main text reports group-level effects, which is an average treatment effect between experimental conditions. Although informative, this analysis can sometimes miss important individual heterogeneity. To provide more details on how each individual performed during this task, here we provide individual results.

Experiment 1.

In addition to group-level effects, we observed individual heterogeneity. Using trials of each participant as the unit of analysis, faces as the independent variable and decisions as the dependent variable, we obtained participants' decision slopes in the initial testing and final testing with 95% confidence intervals. Results showed that whereas 25 out of 58 participants in the fixed sampling condition used positive linear slopes, 34 out of 62 participants in the free sampling condition used positive linear slopes, indicating the persistence of the initial spurious relationship.

In the plots below, the horizontal axis represents each individual participant, and the vertical axis represents the decision slope of that individual. The decision slope is plotted with a point estimate representing the relationship between facial stimuli and donation behaviors, with a 95% confidence interval representing uncertainty. The red dotted line represents the ground truth - which is always zero. Given that participants in Experiments 1 and 2 essentially completed three testing phases - manipulation check, 33 or 15 faces, and 66 or 30 faces, we plotted all three types of decisions - red, green, and blue colors respectively.

Figure S. 4 - 1. Individual slopes in the free sampling condition.

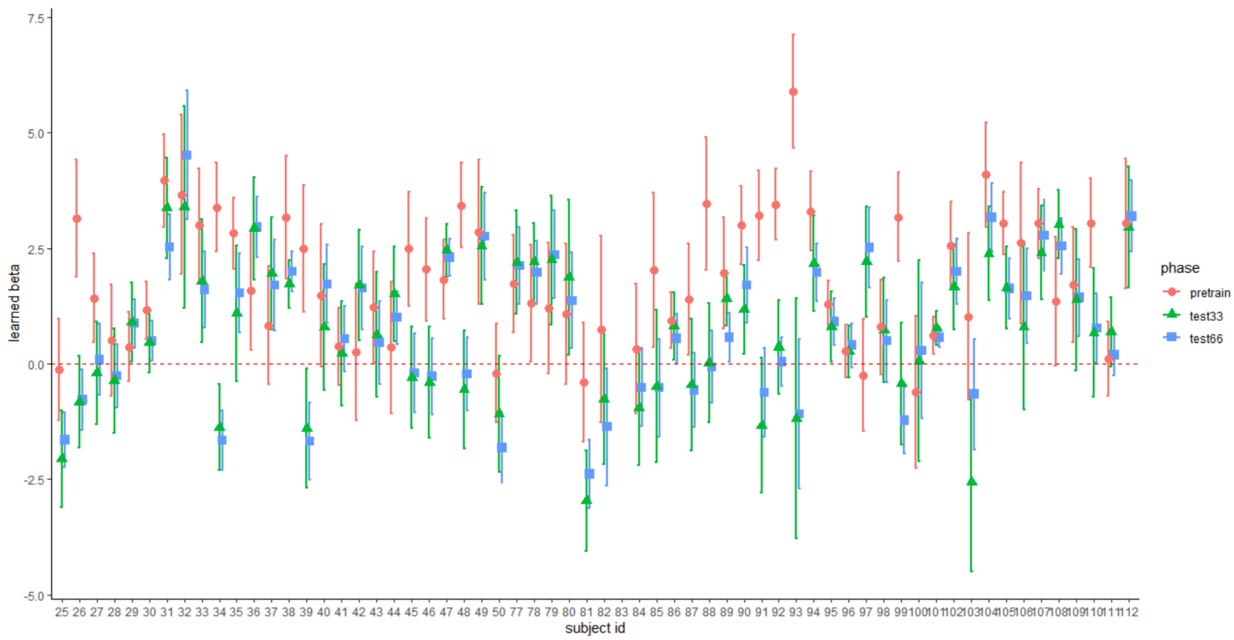
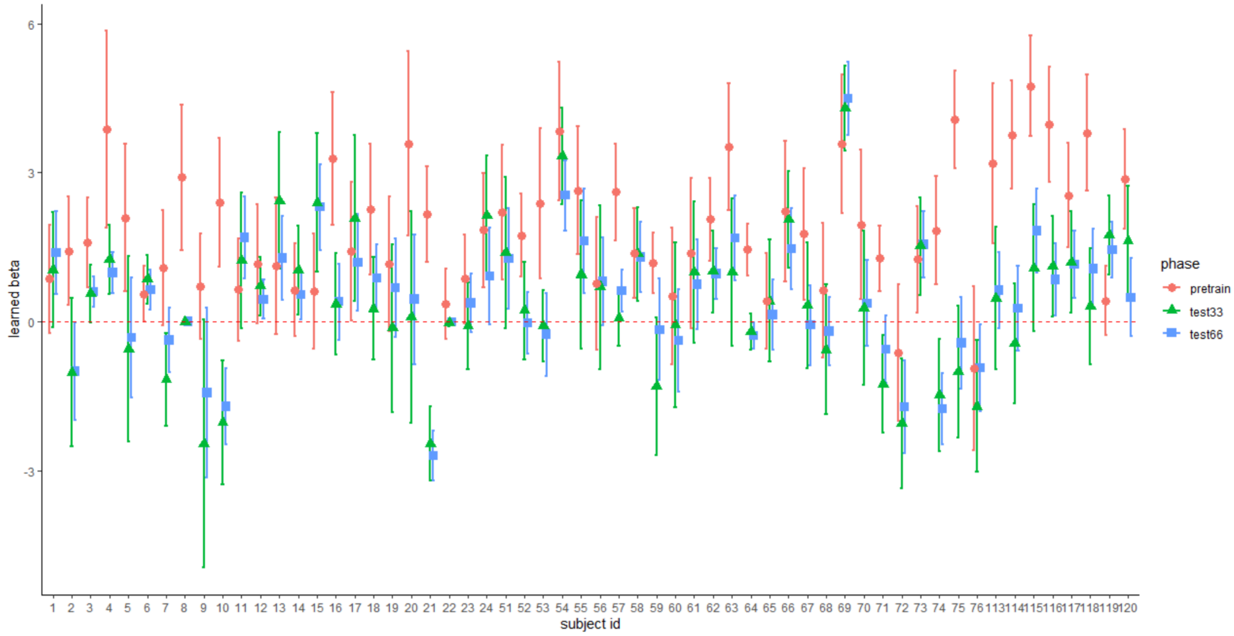


Figure S. 4 - 2. Individual slopes in the fixed sampling condition.



Experiment 2.

Individual-level analysis revealed heterogeneity. Among free sampling participants, 18 out of 55 in the positive condition used statistically significant positive slopes, and 18 out of 52 in the negative condition used statistically significant negative slopes in their final decisions. Whereas among the fixed sampling participants, 15 out of 51 in the positive condition still used positive slopes and 7 out of 48 in the negative condition still used negative slopes in their final decisions. In terms of the zero initial condition, some participants did indeed learn the zero relationships, but around one-third of the participants, 32 out of 114, still used statistically significant linear slopes to make their decisions, which may be consistent with selective attention or selective interpretation mechanisms.

Figure S. 4 - 3. Individual slopes in the free sampling and positive initial condition, ordered by manipulation check sample statistics.

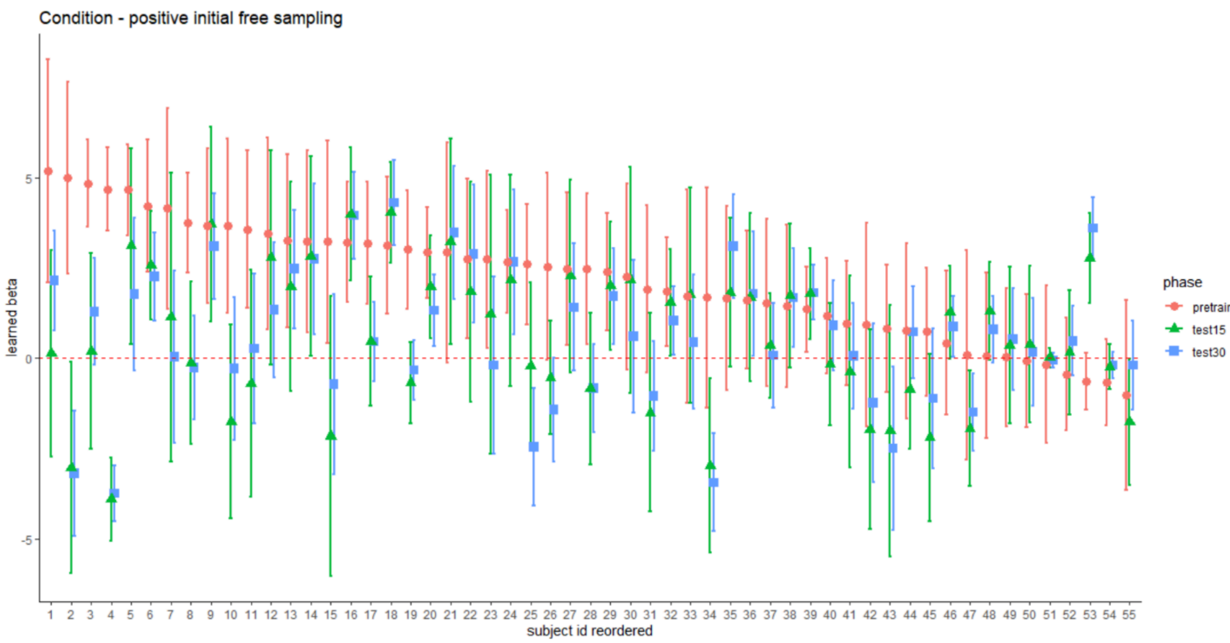


Figure S. 4 - 4. Individual slopes in the free sampling and negative initial condition, ordered by manipulation check sample statistics.

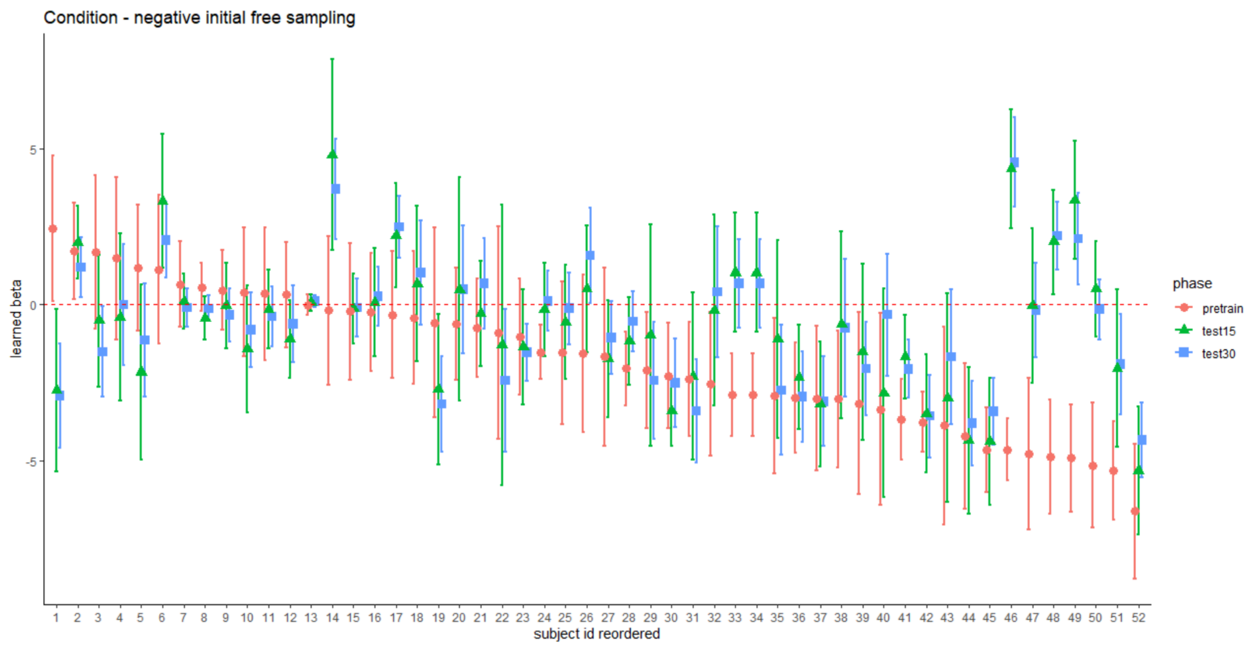


Figure S. 4 - 5. Individual slopes in the free sampling and zero initial condition, ordered by manipulation check sample statistics.

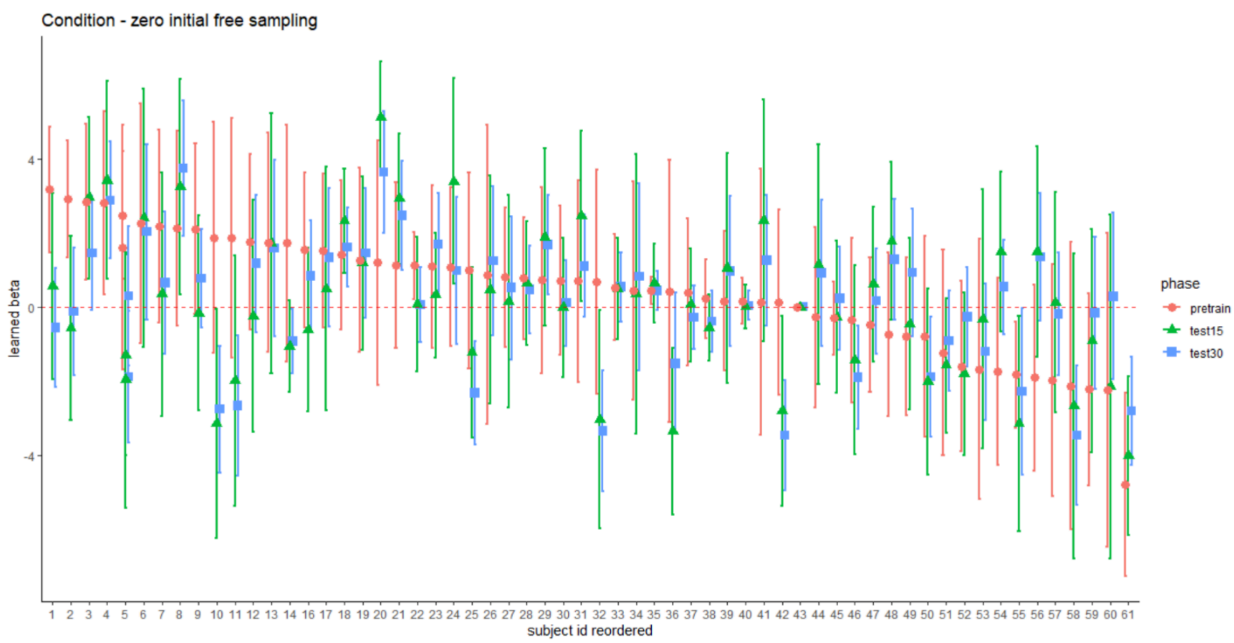


Figure S. 4 - 6. Individual slopes in the fixed sampling and positive initial condition, ordered by manipulation check sample statistics.

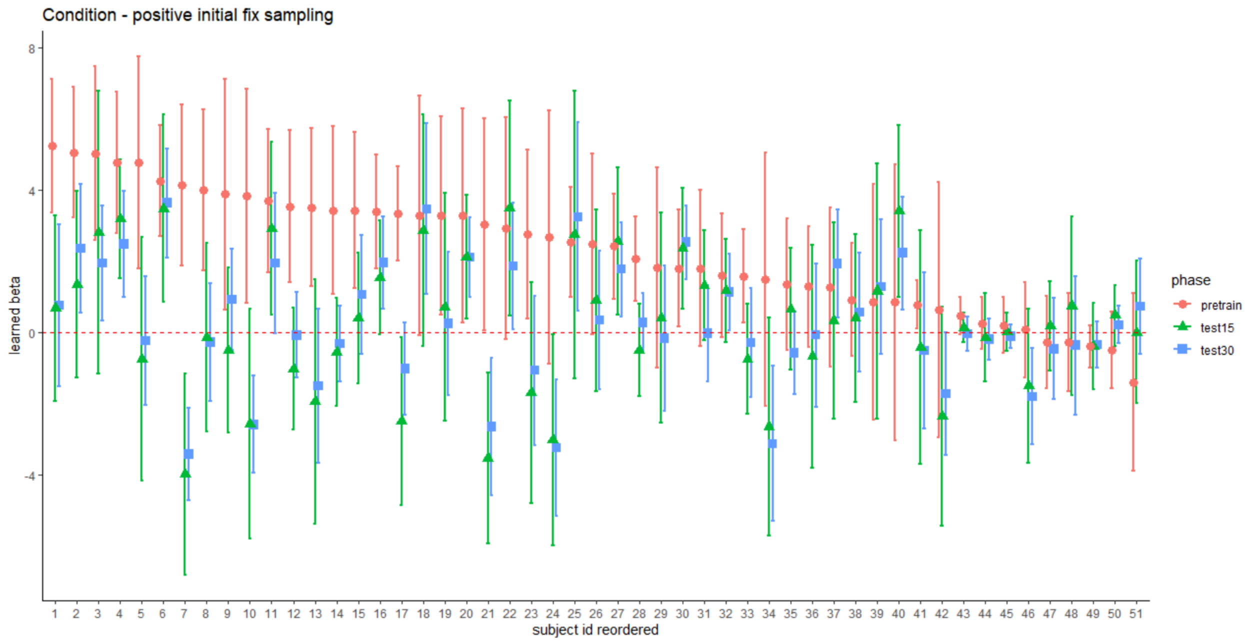


Figure S. 4 - 7. Individual slopes in the fixed sampling and negative initial condition, ordered by manipulation check sample statistics.

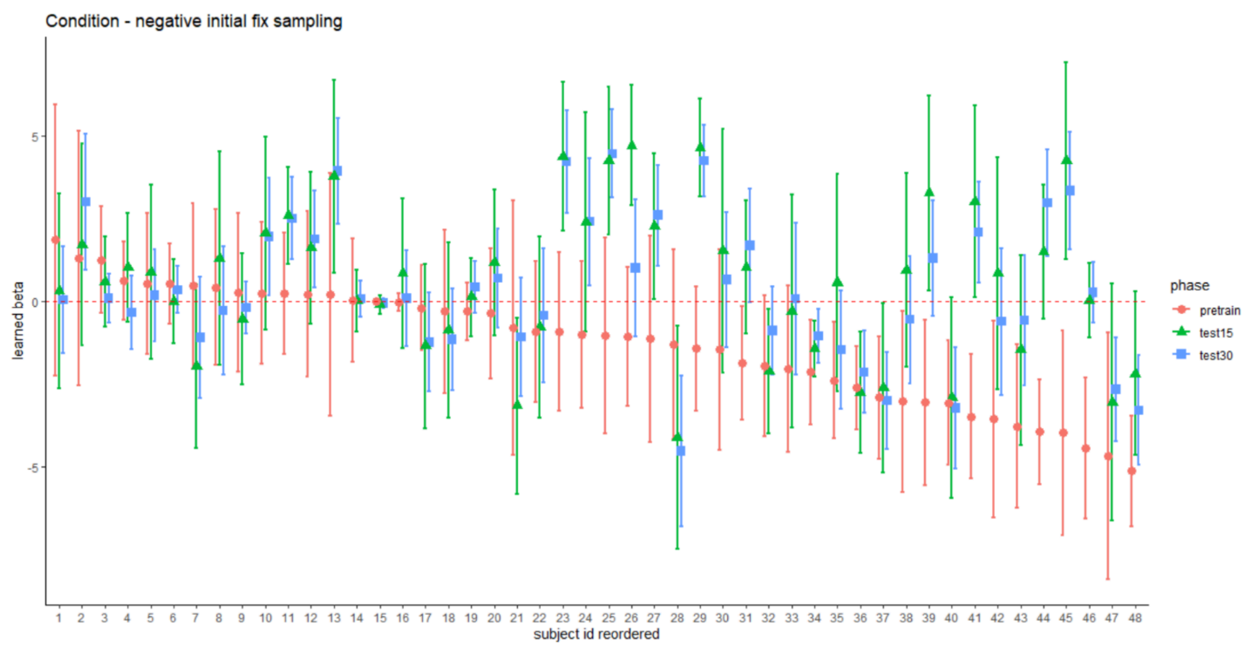
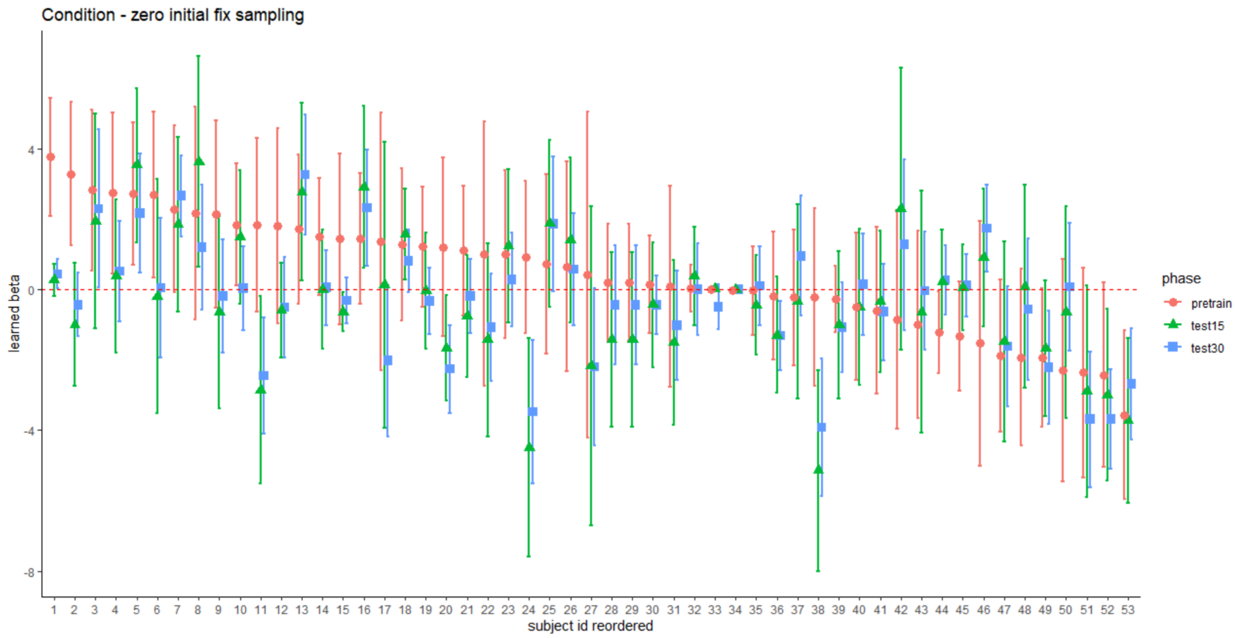


Figure S. 4 - 8. Individual slopes in the fixed sampling and zero initial condition, ordered by manipulation check sample statistics.



Experiment 3.

Individual-level analysis again revealed heterogeneity. 40 out of 95 in the positive initial condition revealed statistically significant positive decision slopes, 26 out of 100 in the negative initial condition revealed statistically significant negative decision slopes, and 28 out of 105 in the zero initial condition revealed decision slopes containing zero.

Figure S. 4 - 9. Individual slopes in the positive initial condition, ordered by decision statistics.

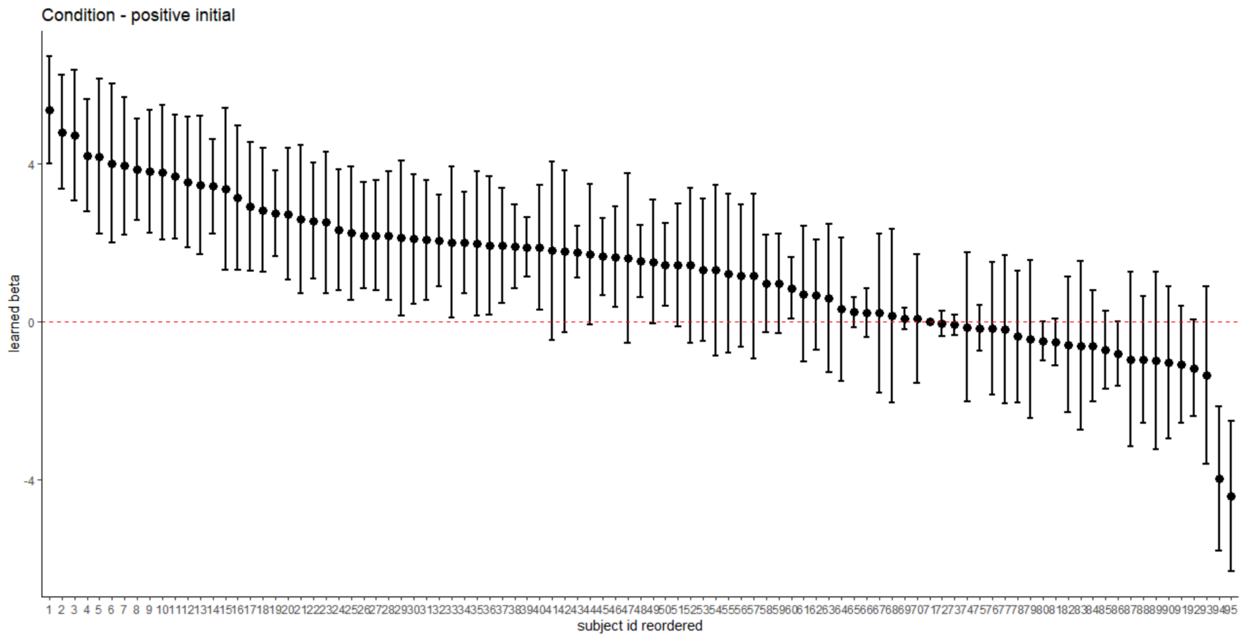


Figure S. 4 - 10. Individual slopes in the negative initial condition, ordered by decision statistics.

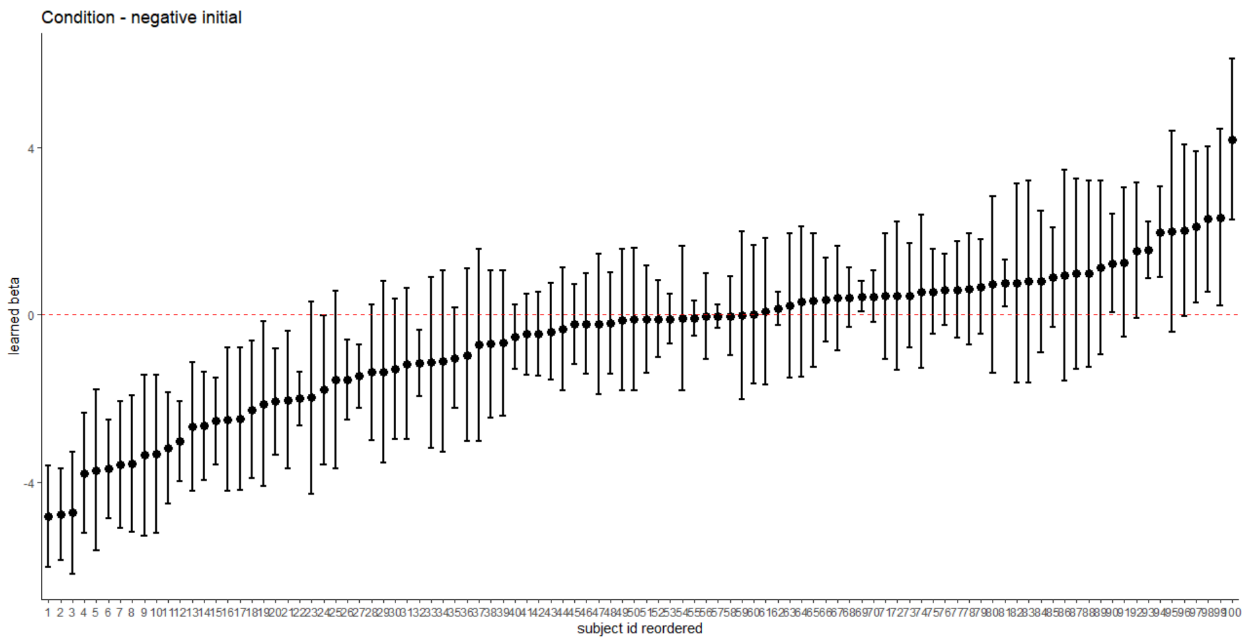
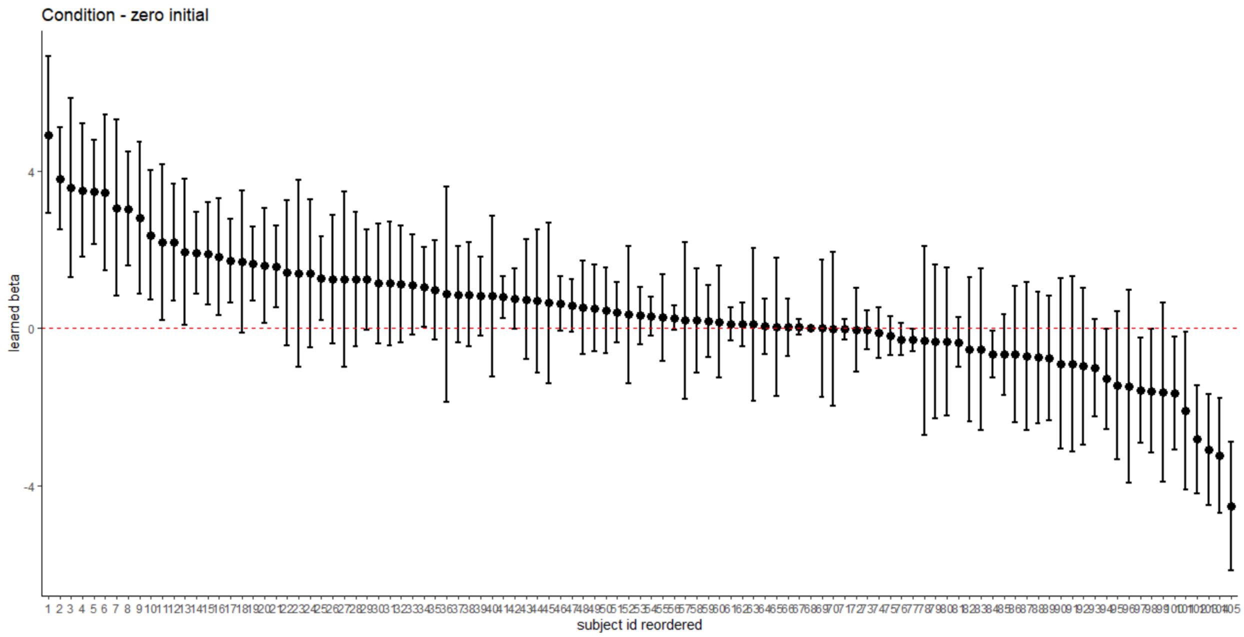


Figure S. 4 - 11. Individual slopes in the zero initial condition, ordered by decision statistics.



This study also examined individual differences in their demographic characteristics and the need for cognitive closure. The analytical model added individual differences as covariates, which did not change the group-level effects significantly. Below we show descriptive statistics between individual characteristics and their decision slopes. The x-axis shows decision slopes (b) and the y-axis shows the count of how many participants used that decision slope(s), grouped by different colors which indicate individual covariates. Consistent with the null effects from the OLS model, there are no obvious individual differences in terms of how participants decide. In other words, quickly learning statistical regularities and insufficiently sampling enough evidence might suggest a general mechanism that is not bounded by individual differences, at least the ones we included in this study.

Figure S. 4 - 12. Gender distribution.

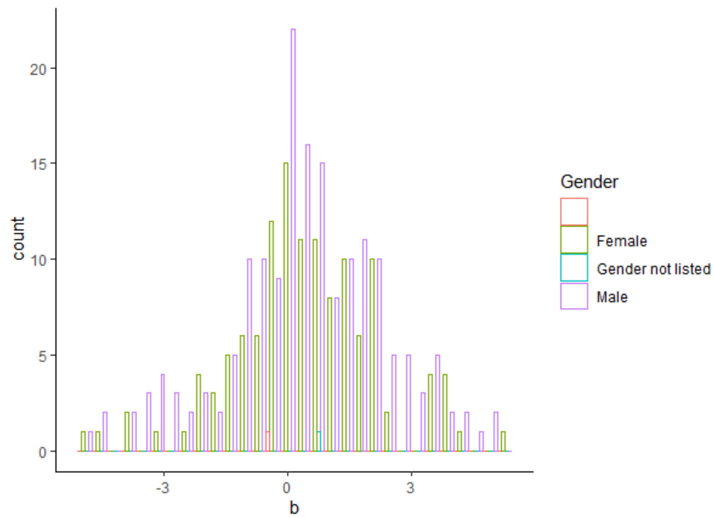


Figure S. 4 - 13. Race distribution.

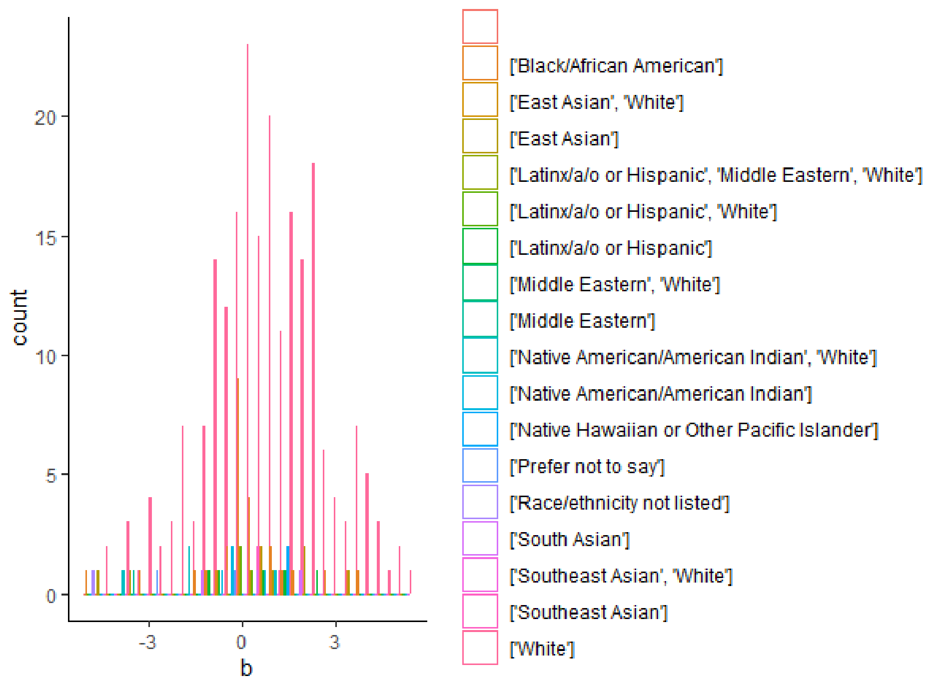


Figure S. 4 - 14. Education distribution.

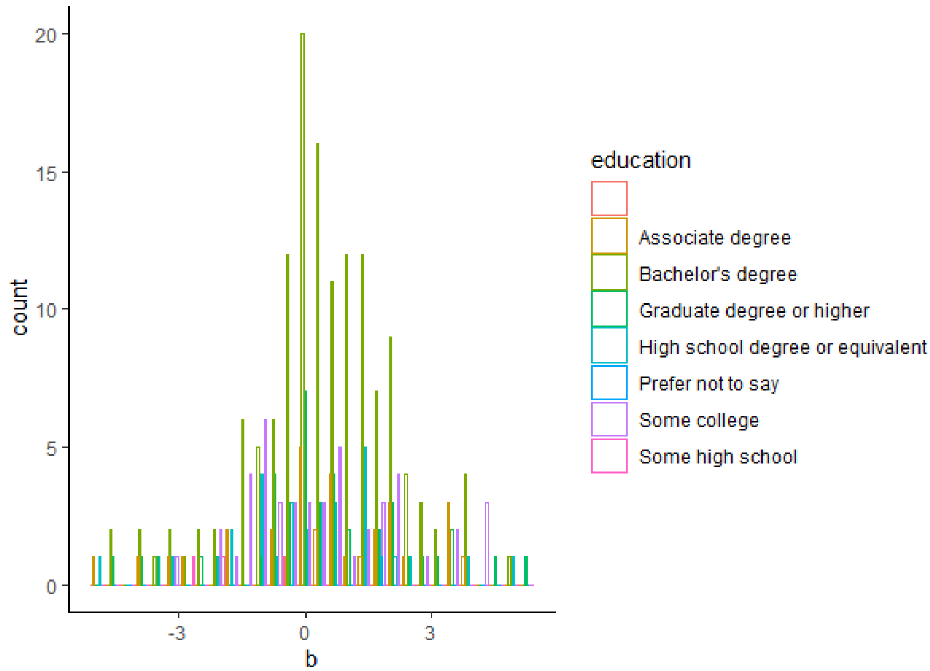
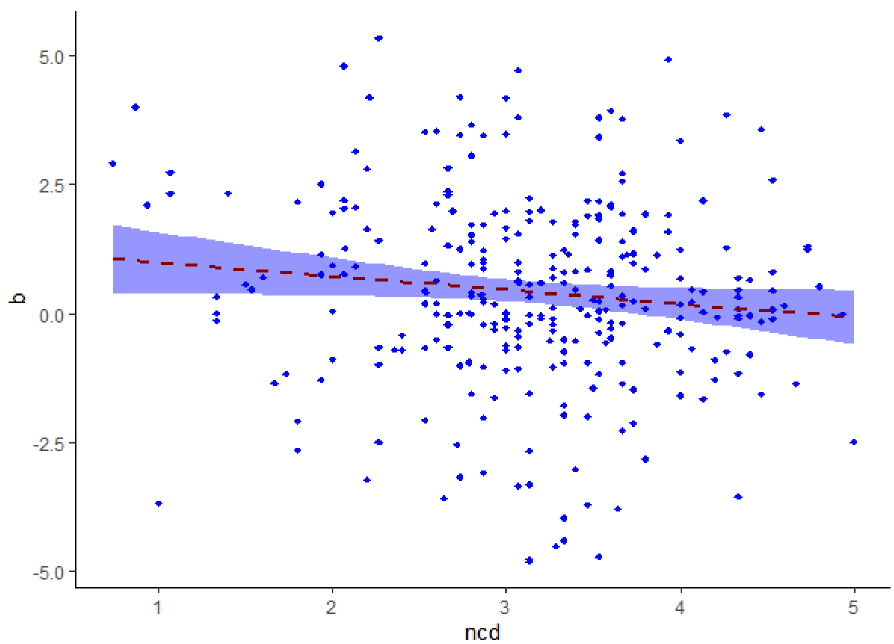


Figure S. 4 - 15. Need for Cognitive Closure distribution.



Experiment 4.

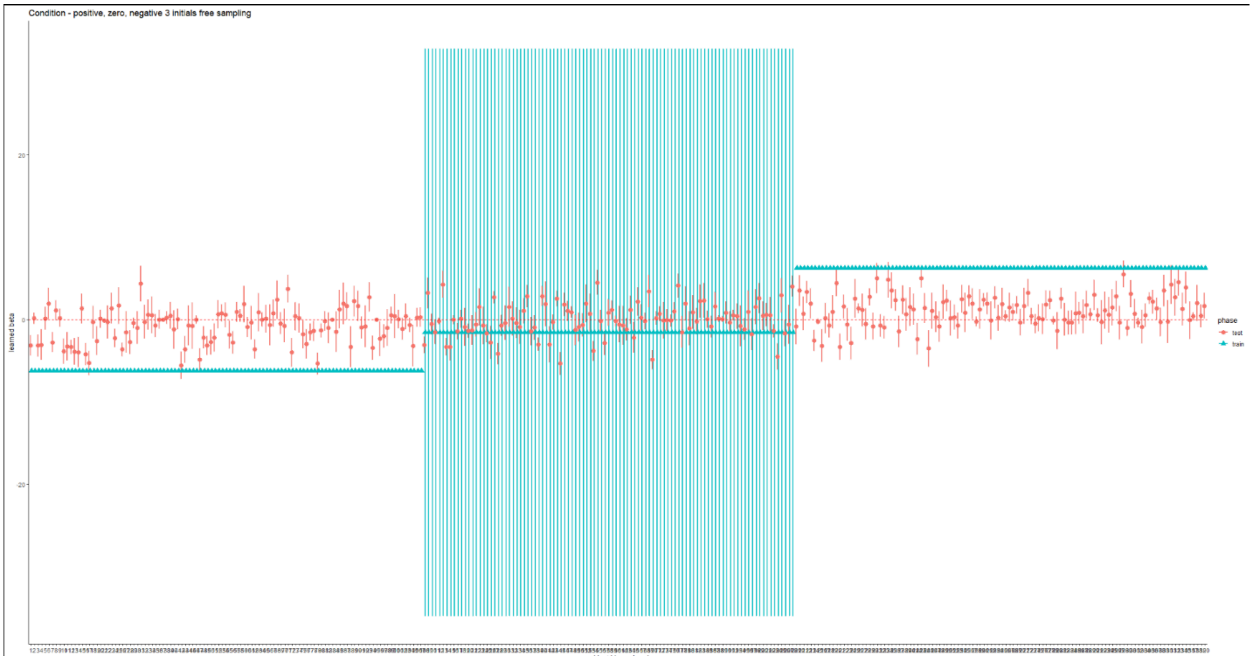
In this study, we added one between-subject condition to explore how categorical information influences this process. This new categorical information cue mimics real-world decisions when people are primed with social categories, regardless of how the categories are related to behaviors. Specifically, we told participants that one of the attributes on which the faces vary is their width. Participants read a hint before entering the game: “Attendees vary by many attributes, the width of their faces is one example.” All else being identical as in the no categorical cue condition. We hypothesized participants who were cued that faces may vary in width could show stronger effects than those in the no cue condition.

Data from the categorical cue condition also replicated Experiment 3: Participants in the positive initial condition ($N = 57$) judged new faces based on a positive linear relationship, $b = .91$, 95% *CI* [.68, 1.13], $p < .001$; participants in the negative initial condition ($N = 53$) used a negative linear relationship, $b = -1.25$, 95% *CI* [-1.49, -1.02], $p < .001$; and participants in the near zero initial condition ($N = 50$) used a slightly negative but closer to zero linear relationship, $b = -.34$, 95% *CI* [-.58, -.10], $p = .005$. Participants in this condition also on average viewed 10 more faces before making decisions.

As expected, the effects were stronger in the categorical cue condition than the no cue condition, but only in the negative and zero initial conditions: A significantly steeper negative trend in the negative (interaction $b = -.79$, 95% *CI* [-1.12, -.47], $p < .001$), and a significant change from slight positive to slight negative in the zero (interaction $b = -.62$, 95% *CI* [-.96, -.29], $p < .001$) initial conditions. There was no difference in the number of samples being viewed between the cue and no cue conditions (interaction $b = -.57$, 95% *CI* [-3.41, 2.27], $p = .694$).

Individual level analysis collapsing across the cue and no cue conditions replicated previous findings: The tendency to persist in using spurious correlations was not related to the need for cognitive closure, total number of trials viewed, total time spent in the task, age, gender, race or levels of education, suggesting that the mechanism is general across demographic groups, at least as measured by our paradigm.

Figure S. 4 - 16. Individual slopes in the positive, negative, and zero initial conditions, with blue dots and lines representing the ground truth and red dots and lines representing participants’ decisions.



Experiment 5.

An individual-level analysis calculated the Pearson correlation between faces and donations in the initial three examples, as well as the Pearson correlation in the 15 testing trials. We hypothesized a positive rank correlation, which was not supported by the data (A Spearman rank-order correlation $r = .04$, $p = .591$). Hence, our individual-level hypothesis was not supported, since participants' final decisions were unrelated to the initial three examples.

However, individual-level results in the participant-generated samples suggest some interesting patterns. We calculated Pearson correlation in participant-generated samples and their testing trials and observed a statistically significant positive rank correlation (A Spearman rank-order correlation $r = .21$, $p = .002$), indicating participants' final decisions were correlated with the relationship they observed in their self-generated sample.

Figure S. 4 - 17. Individual correlations between face stimuli and behavioral indicators from the initial 3 examples and the testing 15 trials. The red dotted line is the ground truth correlation, red dots represent participants' decisions whereas blue triangles represent the initial sample statistics. Again, results are ordered by initial sample statistics.

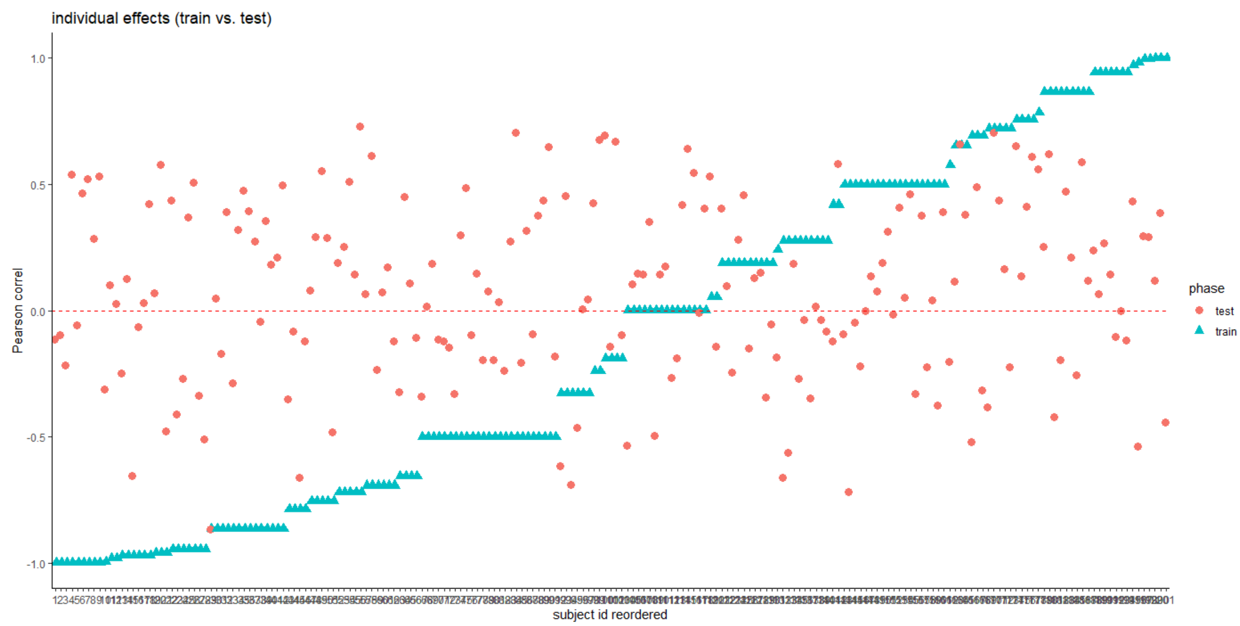


Figure S. 4 - 18. Individual correlations between face stimuli and behavioral indicators from the participant-self-generated examples and the testing 15 trials.

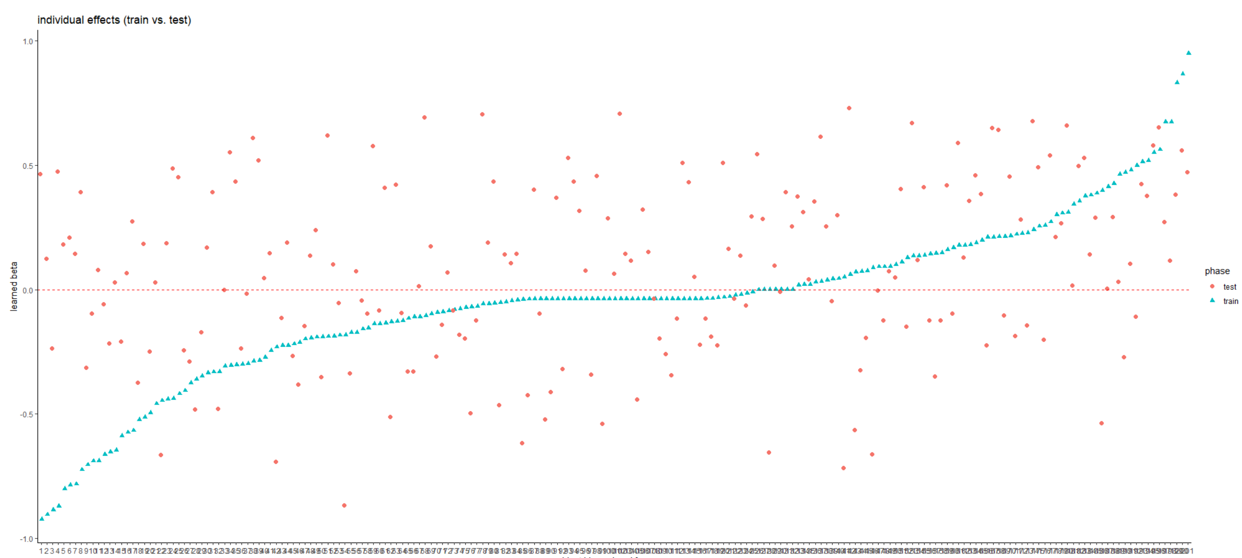
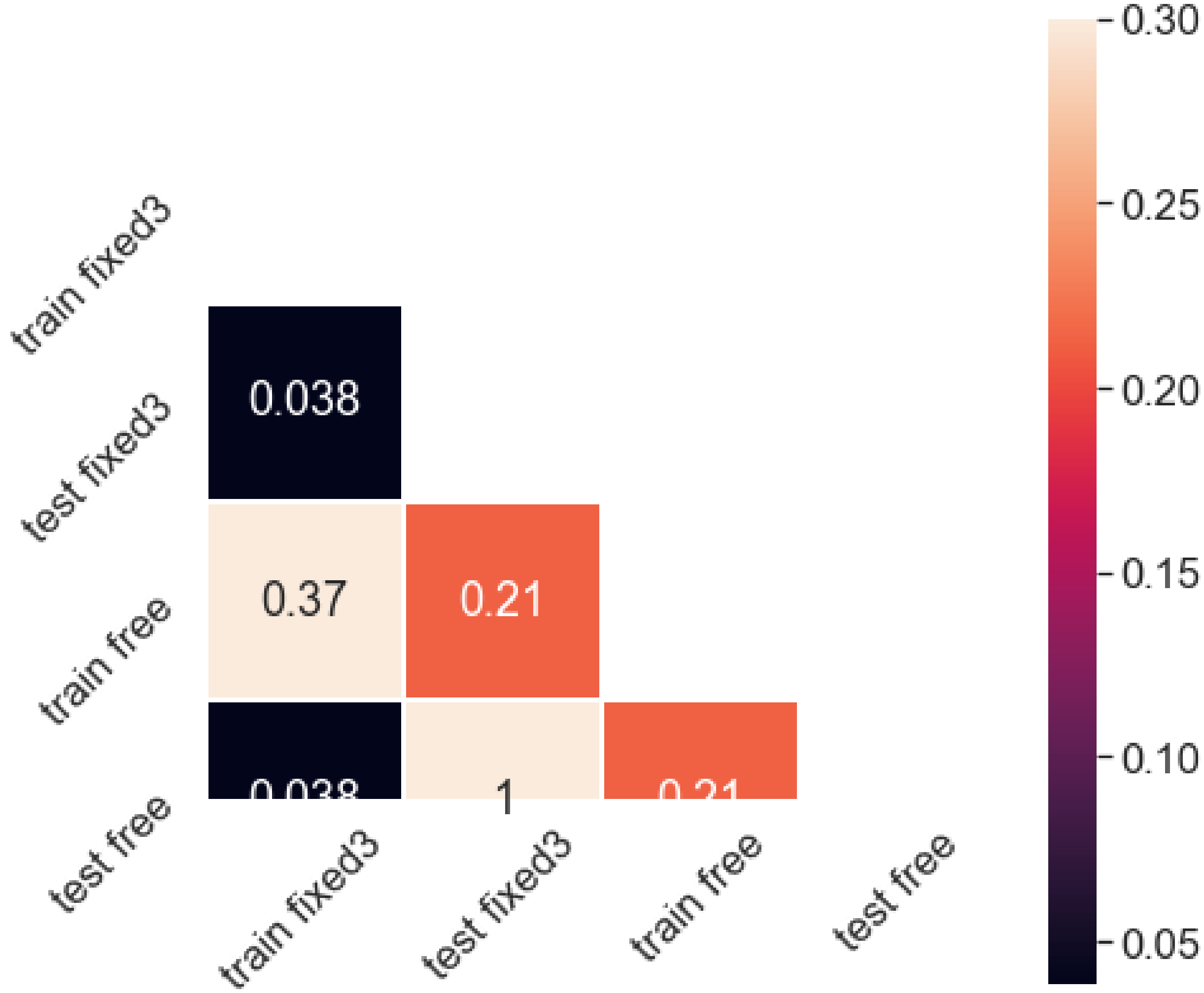


Figure S. 4 - 19. Individual correlations in fixed three initial samples, participant-generated samples, and testing trials, were demonstrated via Spearman rank correlations.



Experiment 6.

Below we visualize results from the three sub-hypotheses in the main text and in the pre-registration report.

Figure S. 4 - 20. Number of hints being asked in conditions where the initial three examples show positive, negative, and zero correlations.

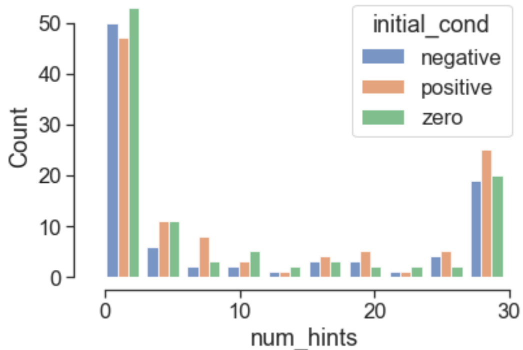


Figure S. 4 - 21. Correlations between revealed hints and participants' final decisions in three respective conditions.

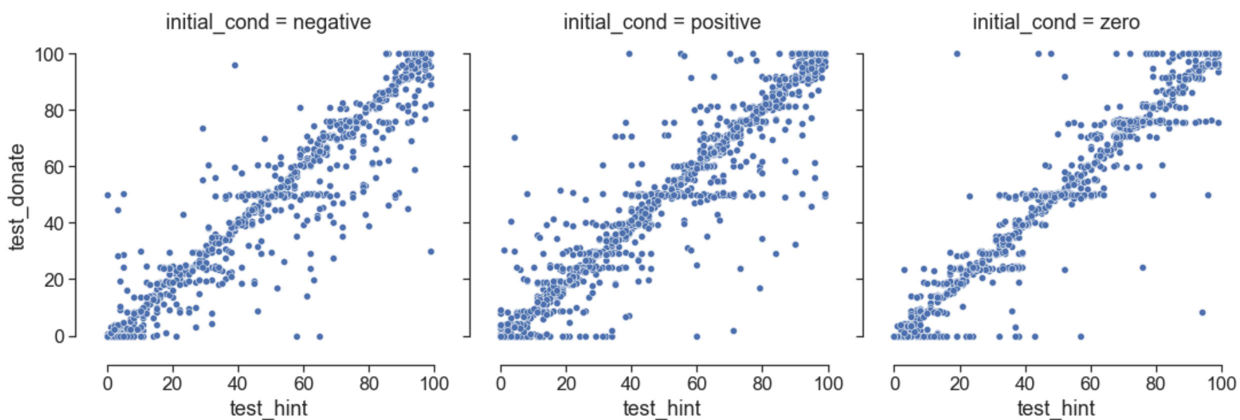


Figure S. 4 - 22. Final decisions were made in conditions where participants asked for fewer hints (left; fewer than 10 pieces), more hints (middle; more than 20 pieces), and in-between.