# Electronic Supplementary Materials

Abstract

Electronic Supplementary Materials for "Who knows what? Bayesian Competence Inference guides Knowledge Attribution and Information Search"

**Contents**

# 1 Materials

## 1.1 Theme 1: Solar system

1. What is the name of the first planet of the solar system? Mercury
2. What is the name of the first man that stepped on the Moon? Neil Armstrong
3. Which planet in the solar system is the only one that rotates clockwise? Venus
4. Which planet has a hexagon shaped cloud formation on its north pole? Saturn
5. What is the orbital period of the Earth around the Sun in days? 365
6. What are the only two planets of our solar system that do not possess any moon? Mercury and Venus
7. What is Earth's only natural satellite? The Moon
8. What is the only liveable planet of our solar system? Earth
9. How old is the Sun (rounded up to the nearest billion years)? 5
10. What is the name of the first human object that crossed the limits of our solar system? Voyager 1
11. What is the last planet of our solar system? Neptune
12. What is the largest celestial object of the asteroid belt? Ceres
13. Pluto used to be the 9th planet of our solar system, but is no longer considered as a planet by scientists nowadays but as a . . . ? Dwarf planet
14. Between which planets of our solar system is located the asteroid belt? Mars and Jupiter
15. What is the name of the circumstellar disc located beyond the orbit of Neptune? Kuiper belt

## 1.2 Theme 2: American history

16. When did Christopher Columbus discover the Americas? 1492
17. What is the date of the United States Declaration of Independence? 1776
18. What is the year of the beginning of the American Civil War? 1861
19. What is the year of the beginning of the Confederation Period? 1783
20. What is the date of foundation of the city of Los Angeles? 1781
21. When did the US declare war on Japan? 1941
22. Independence Day was first established as a holiday by Congress in what year? 1870
23. Who was the first President to live in the White House? John Adams
24. Where was the first Fourth of July Celebration with a firework display held? Boston
25. Which President was in office between 1893 and 1897? Grover Cleveland
26. When did Franklin D. Roosevelt die? 1945
27. Which was the first state to announce its secession from the Union before the Civil War began? South Carolina
28. When did prohibition start and end in the United States? 1920-1933
29. What is the name of the document with the first ten amendments to the Constitution that detail the protection of individual liberties? The Bill of Rights
30. What was called the first American Constitution? The Articles of Confederation

### 1.3 Theme 3: Superheroes

31. What is the name of Batman's mother? Martha
32. What weapon does Thor use? Hammer
33. What is the real name of Spider-Man? Peter Parker
34. Who is the Human Torch to the Invisible Woman? Brother
35. Who created the character of Iron Man? Stan Lee
36. Batman protects which city? Gotham
37. Which superhero has a magic lasso and bullet-proof bracelets? Wonder Woman
38. The Green Lantern gains his power from which object? A ring
39. Wonder Woman comes from which island? Paradise Island
40. Which newspaper does Spiderman work for? The Daily Bugle
41. What is the name of the archnemesis of the Fantastic Four? Dr. Doom
42. What is the name of the archenemy of Aquaman? Malefic
43. Who killed Superman in the 1993 comic? Doomsday
44. Who was Aquaman's sidekick? Aqualad
45. What is the family name of Alfred, Batman's butler? Pennyworth

## 2 Stan Model Exploratory Analyses and Mathematical Derivations

The Bayesian ideal observer model assumes that participants infer an agent's competence $\theta$ from observing their performance on questions of varying difficulty. We first estimated this model via MLE to find the best-fitting parameters (using a grid over competence). We present now present a robustness analysis by fitting the same model in Stan with Markov chain Monte Carlo (specifically the No-U-Turn sampler, NUTS). The main results are reported in the manuscript; here we detail the closed-form expressions used in the Stan implementation.

### 2.1 Computational model

**2.1.1 Marginal probability of success at a given difficulty.** As described in the main manuscript, the Bayesian model assumes that an individual's latent competence $\theta$ varies across individuals following a normal distribution:

$$\theta \sim N(\mu, \sigma^2),$$

The probability of solving a question of difficulty $\beta$, conditional on $\theta$ is given by:

$$\Pr(S = 1 \mid \theta, \beta) = \Phi\left(\frac{\theta - \beta}{\varepsilon}\right),$$

Where $\Phi$ is the standard normal CDF.

To compute the marginal probability of success, we want $P(S = 1|\beta)$ integrating out $\theta$. We previously used a grid approximation in MLE but such method is extremely slow in Stan, we therefore fitted the model using a closed-form solution of our model that we describe below.

We set $\theta = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0, 1)$. Then

$$\Pr(S = 1 \mid \beta) \;=\; \mathbb{E}_Z\left[\Phi\left(\frac{\mu - \beta}{\varepsilon} + \frac{\sigma}{\varepsilon} Z\right)\right].$$

We now use the standard probit-normal convolution identity (see e.g Rasmussen & Williams, 2006). For $Z \sim \mathcal{N}(0, 1)$ and constants $a, b$, we have:

$$\mathbb{E}\big[\Phi(a + bZ)\big] \;=\; \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right)$$

Applying the identity with $a = (\mu - \beta)/\varepsilon$ and $b = \sigma/\varepsilon$ gives the closed form solution for the marginal probability of success:

$$\boxed{P(S = 1 \mid \beta) \;=\; \Phi\left(\frac{\mu - \beta}{\sqrt{\sigma^2 + \varepsilon^2}}\right)}$$

For convenience, we define:

$$\sigma_* \equiv \sqrt{\sigma^2 + \varepsilon^2}, \qquad x(\beta) \equiv \frac{\mu - \beta}{\sigma_*}.$$

Then $P(S = 1 \mid \beta) = \Phi\big(x(\beta)\big)$.

**2.1.2 Joint probability of two answers at two difficulties.** The Bayesian model needs to predict the probability of $S_{new} = 1$ given $\beta_{new}$ after observing $S_{\text{obs}}$ on difficulty $\beta_{\text{obs}}$.

Conditional on $\theta$, we can define the joint probability of having two correct answers with:

$$P(S_{\text{new}} = 1, S_{\text{obs}} = 1 \mid \theta, \beta_{new}, \beta_{obs}) = \Phi\Big(\tfrac{\theta - \beta_{\text{new}}}{\varepsilon}\Big)\, \Phi\Big(\tfrac{\theta - \beta_{\text{obs}}}{\varepsilon}\Big).$$

Taking expectations over $\theta$ yields the following joint probability which is a bivariate probit:

$$p_{\text{joint}} = \mathbb{E}_\theta\Big[\Phi\Big(\tfrac{\theta - \beta_{\text{new}}}{\varepsilon}\Big)\, \Phi\Big(\tfrac{\theta - \beta_{\text{obs}}}{\varepsilon}\Big)\Big] = \Phi_2\big(x_{\text{new}}, x_{\text{obs}}; \rho\big),$$

where $\Phi_2$ is the standard bivariate normal CDF, with

$$x_{\text{obs}} = x(\beta_{\text{obs}}), \qquad x_{\text{new}} = x(\beta_{\text{new}}), \qquad \rho = \frac{\sigma^2}{\sigma^2 + \varepsilon^2}.$$

**2.1.3 Close-form solution.** Let $p_{\text{obs}} = P(S = 1 \mid \beta) = \Phi(x_{\text{obs}})$. Then

$$P(S_{\text{new}} = 1 \mid S_{\text{obs}} = 1) = \frac{p_{\text{joint}}}{p_{\text{obs}}}, \qquad P(S_{\text{new}} = 1 \mid S_{\text{obs}} = 0) = \frac{\Phi(x_{\text{new}}) - p_{\text{joint}}}{1 - p_{\text{obs}}}.$$

## 2.2 Results

The main results are reported in the manuscript. We report in the following tables additional fit statistics. Diagnostics for all fits in Study 1 and 2 indicated reliable estimates (all Pareto-$k < 0.7$).

Table S1
*Posterior summaries and PSIS-LOO diagnostics for Study 1 Stan models*

| Quantity | Bayesian | Threshold |
|---|---|---|
| Mu | 0.01 (0.02) | NA |
| Sigma | 0.72 (0.03) | NA |
| Noise | 0.67 (0.03) | NA |
| Tau | 0.42 (0.01) | 1.32 (0.02) |
| | | |
| ELPD | -33566.79 (109.30) | -38091.65 (74.85) |
| p_LOO | 4.01 | 0.92 |
| LOOIC | 67133.59 (218.60) | 76183.29 (149.70) |
| Delta ELPD | 0.00 (0.00) | -4524.85 (89.21) |

Table S2
*Posterior summaries and PSIS-LOO diagnostics for Study 2 Stan models*

| Quantity | Bayesian | Threshold |
|---|---|---|
| Mu | 0.32 (0.01) | NA |
| Sigma | 0.60 (0.01) | NA |
| Noise | 0.69 (0.03) | NA |
| Tau | 0.45 (0.01) | 1.41 (0.02) |
| | | |
| ELPD | -30928.44 (103.16) | -34969.36 (67.86) |
| p_LOO | 4.05 | 0.97 |
| LOOIC | 61856.87 (206.31) | 69938.73 (135.72) |
| Delta ELPD | 0.00 (0.00) | -4040.93 (84.66) |

## 3 Individual Analyses

In this appendix, we provide additional analyses on individual level behaviours. In the main manuscript, we reported that some participants are better predicted by heuristics model than by our main Bayesian model. Here, we examine if this categorization can be explained by significant differences in the estimation of (a) difficulty and (b) predictive accuracy. We show that in both Study 1 and 2, participants not categorized by the Bayesian model are worse in predicting the true difficulty of questions and are worse in predicting if the evaluated individual got the question right or wrong.

### 3.1 Study 1

### 3.2 Study 2

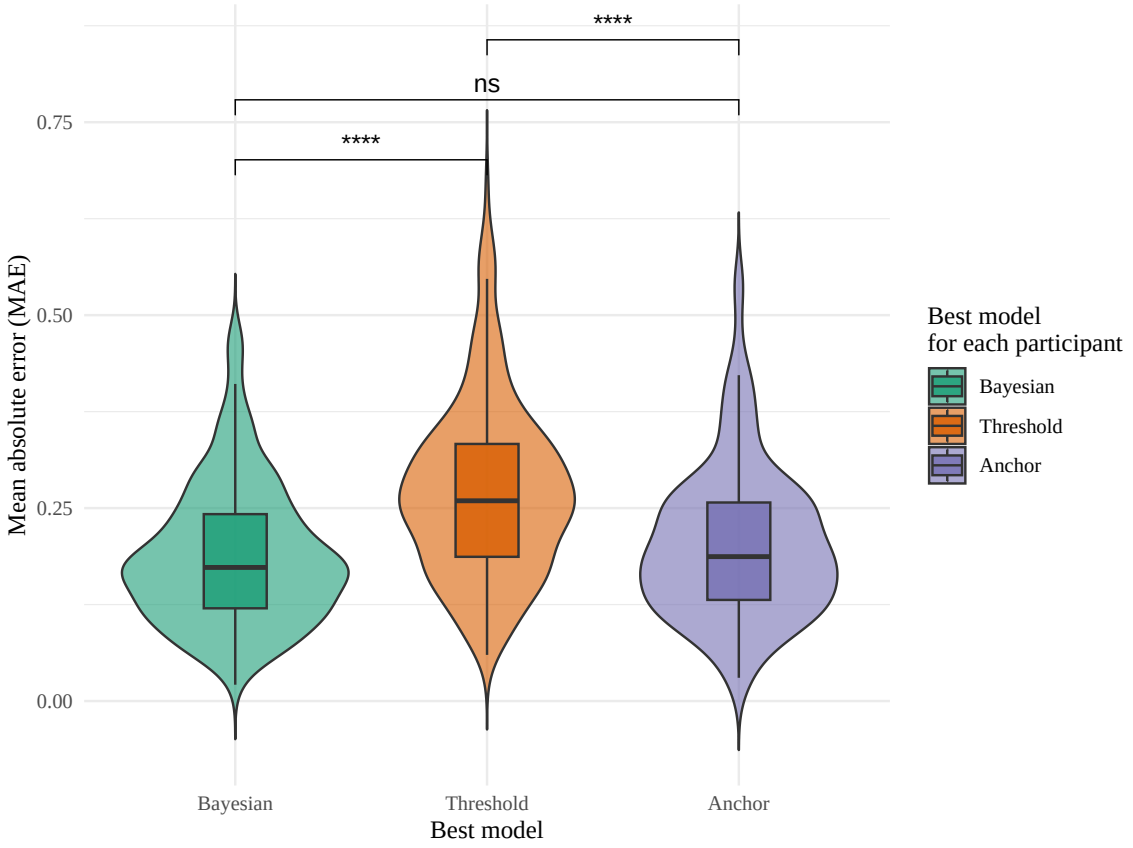


*Figure S1*. Study 1: Per-participant mean absolute error for predicting difficulty, grouped by best-fitting model.

## 4 Robustness check with true difficulty

In the main analyses, we used participants' subjective judgments of question difficulty. Here, we present a robustness check using the "true" difficulty of questions as estimated by an Item Response Theory (IRT) Rasch model fitted to participants' actual performance. This allows us to test whether our findings hold when using an objective measure of difficulty rather than subjective judgments.

### 4.1 Study 1

We refit all models using the IRT-estimated question difficulties instead of judged difficulties. The model fitting procedure remained identical to the main analysis.

The Bayesian model continues to outperform both heuristic models when using true difficulty estimates. The BIC for the Bayesian model (68537) remains substantially lower than both the Threshold heuristic (71540) and the Anchor heuristic (74481). The BIC difference between the Bayesian model and the next best model is 3003, providing strong evidence in favor of the Bayesian model.

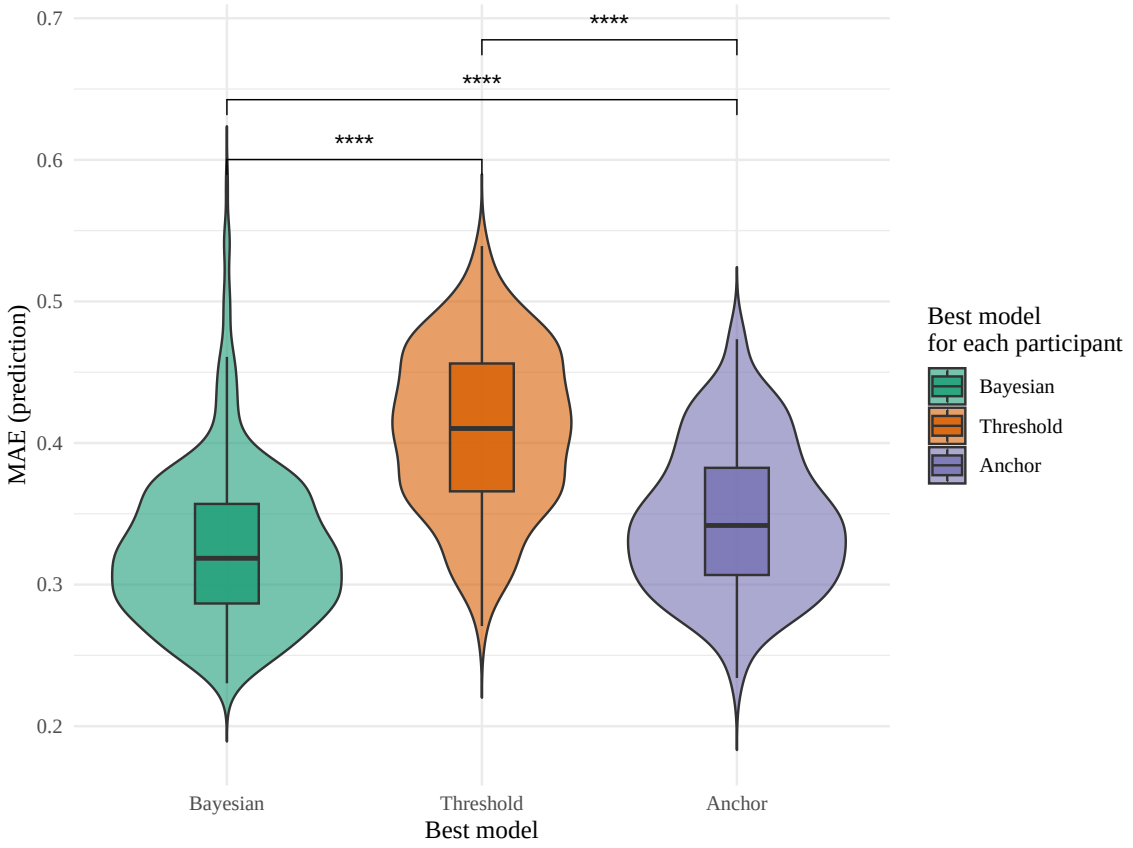


*Figure S2*. Study 1: Per-participant mean absolute error for predictions (Yes/No) vs true conditional probability, grouped by best-fitting model.

Table S3
*Study 1: Model comparison using IRT-estimated true difficulty*

| Model | LLMax | BIC | Mu | Sigma | Noise | Tau | Difference |
|---|---|---|---|---|---|---|---|
| Bayesian | -34246 | 68537 | 1.5 | 8 | 2.12 | 0.44 | |
| Threshold | -35764 | 71540 | | | | 0.86 | |
| Anchor | -37230 | 74481 | | | | 1.33 | 1.7 |

### 4.2 Study 2

Similarly to Study 1, we refit all models using IRT-estimated question difficulties.

As with Study 1, the Bayesian model remains more predictive of participants' behaviors over both heuristic models. The BIC for the Bayesian model (62607) is lower than the Threshold heuristic (65802) and the Anchor heuristic (67455). The BIC difference of 3195 between the Bayesian model and the next best alternative continues to provide strong evidence for the Bayesian account.

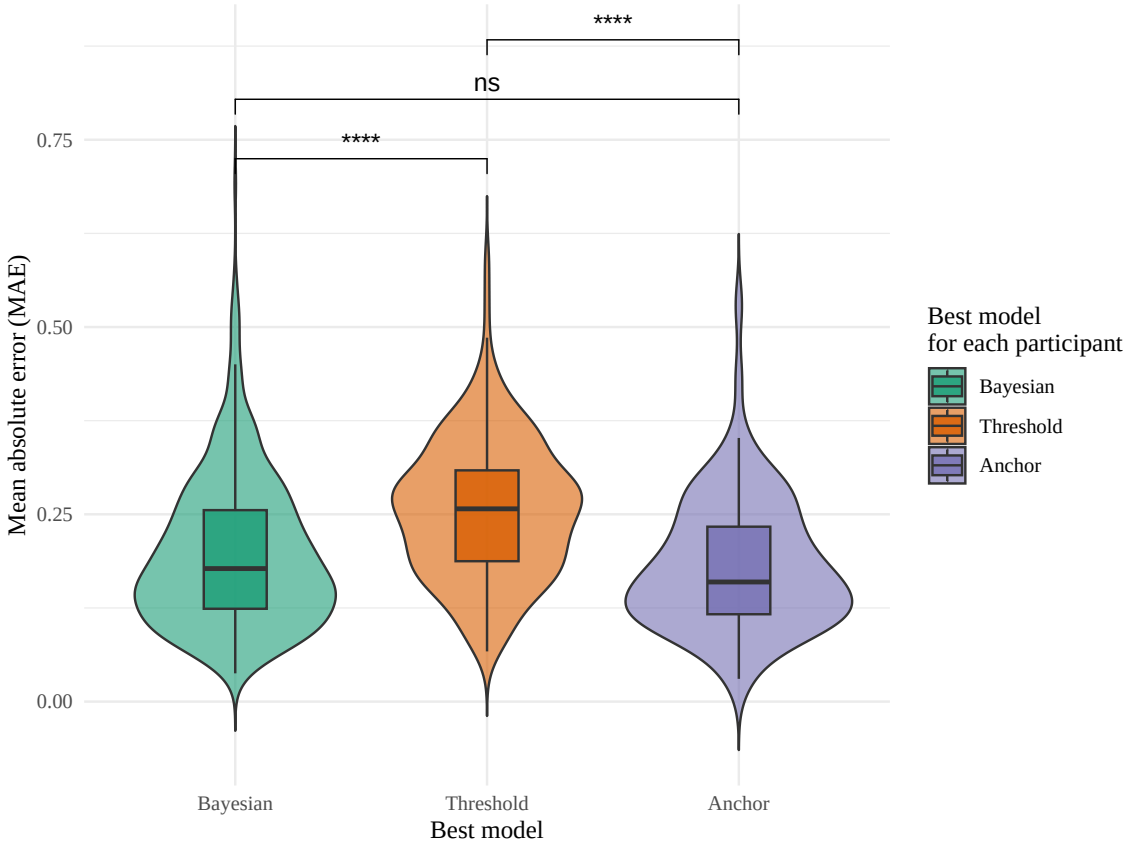


*Figure S3*. Study 2: Per-participant mean absolute error in prediction of difficulty, grouped by best fitting model

Table S4
*Study 2: Model comparison using IRT-estimated true difficulty*

| Model | LLMax | BIC | Mu | Sigma | Noise | Tau | Difference |
|---|---|---|---|---|---|---|---|
| Bayesian | -31282 | 62607 | 1.73 | 2.53 | 2.23 | 0.46 | |
| Threshold | -32896 | 65802 | | | | 0.91 | |
| Anchor | -33717 | 67455 | | | | 1.30 | 1.61 |

## 5 Inference of difficulty

In Study 2, participants estimated the difficulty of the observed question. Figure S5 shows the relationship between true difficulty (as estimated from Study 1) and participants' estimated difficulty, split by whether they observed a success or failure on that question.

To quantify how much variance in difficulty judgments is associated with observing a success/failure versus stable item differences, we fit two linear models with the estimated difficulty as the dependent variable. With items included as fixed effects, the item-only model achieved $R^2 = 0.265$, and adding success/failure increased $R^2$ to 0.269 ($\Delta R^2 = 0.004$).

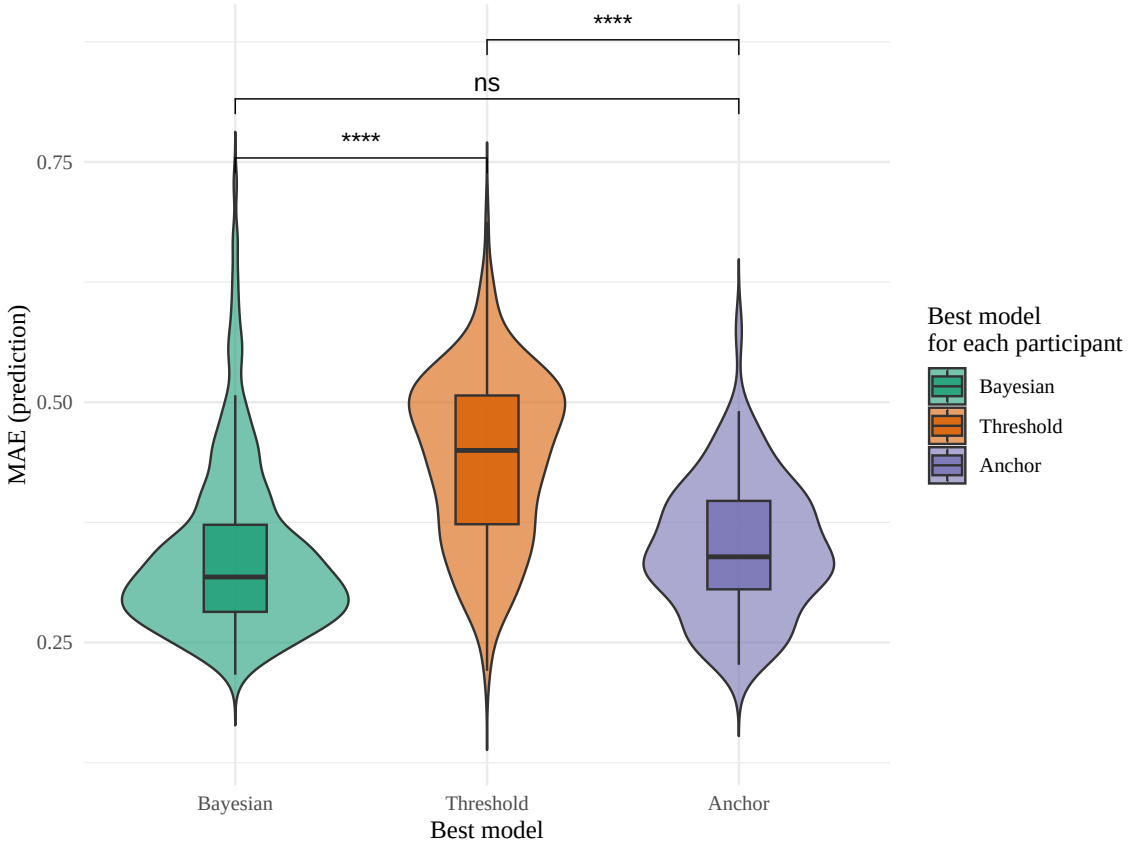


*Figure S4*. Study 2: Per-participant mean absolute error for predictions (Yes/No) vs true conditional probability, grouped by best-fitting model.

## 6 References

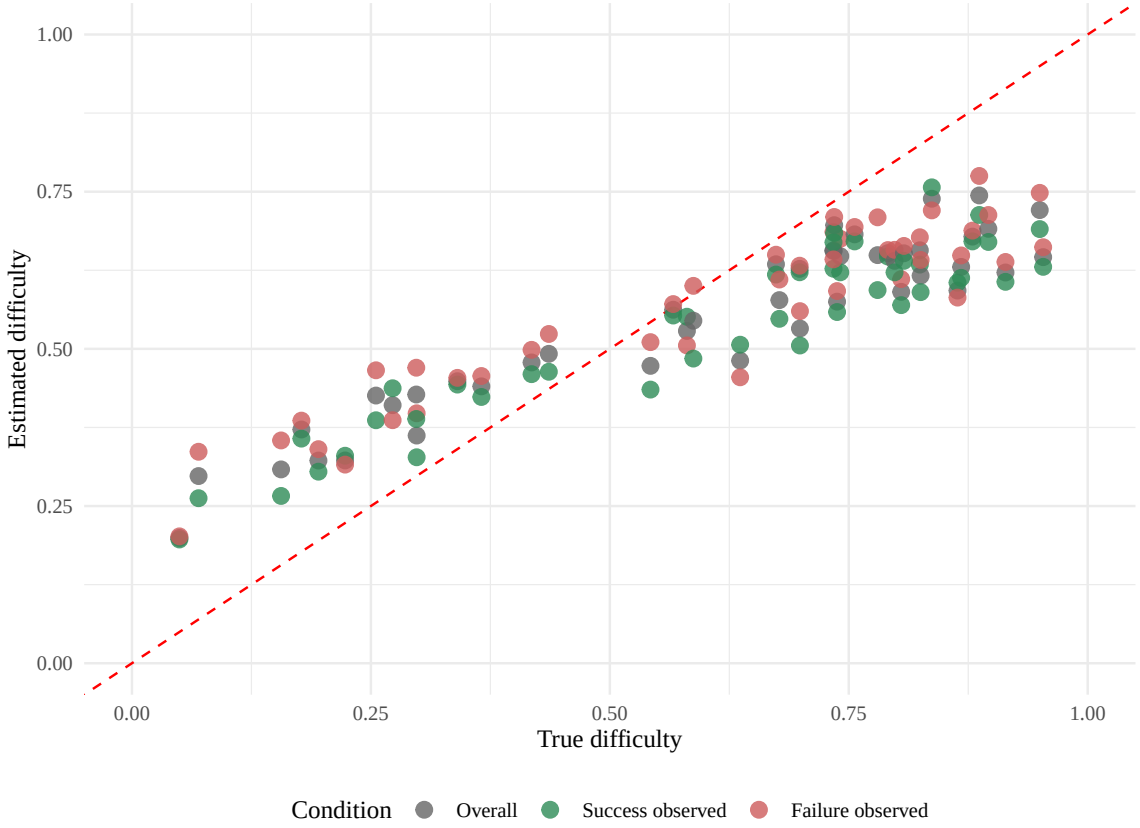


*Figure S5*. Estimated difficulty as a function of true difficulty in Study 2. Each question is represented by three data points: the overall average estimation (grey), the average estimation when a success was observed (green), and the average estimation when a failure was observed (red). The dashed line represents perfect accuracy.