Who knows what? Bayesian Competence Inference guides Knowledge Attribution and Information Search

Marius Mercier[1], Olivier Morin[†,1], Hugo Mercier[†,1], & Tadeg Quillien[†,2]

[1] Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France
[2] Department of Psychology, University of Edinburgh, United Kingdom

This manuscript is accepted for publication at Cognition.

Abstract

One of the main challenges of social cognition is inferring the competence of others, which often occurs in contexts of limited information. Recent research suggests that people can successfully infer the competence of others through Bayesian inference, but prior studies have relied largely on artificial stimuli and competence benchmarks, leaving it unclear whether these rational principles generalize to naturalistic settings with rich prior knowledge. Using trivia questionnaires, we test whether people can infer others' competence and search for informative evidence in a near-optimal way, consistent with rational Bayesian principles. In Studies 1 and 2, participants were presented with an individual's performance on a trivia question and predicted the individual's ability to answer other trivia questions from the same theme. Replicating and extending past results, we observe that participants accurately predict performance from limited information. Computational modelling shows that participants' inferences are better described by Bayesian processes than by plausible heuristics, suggesting that participants rationally integrate new information with their prior expectations about others' competence. Study 3 shows that participants can select which information would be most diagnostic for inferring an individual's competence, again in a manner consistent with Bayesian rationality. Overall, our results suggest that people approximate a rational Bayesian model both when searching for and when integrating information about others' competence.

Keywords: Computational Modeling; Social Cognition; Competence; Knowledge Attribution; Information Search

1 Introduction

Estimating an interlocutor's knowledge and skills is of tremendous importance: judgements of competence influence who we learn from (Birch et al., 2008; Harris et al., 2018; Laland, 2004; Lane et al., 2013; Mercier, 2020; Najar et al., 2020; Sperber et al., 2010), who we cooperate with (Cuddy et al., 2007; Magid et al., 2018; Xiang et al., 2023), who we hire (Cuddy et al., 2011; Fousiani et al., 2023; Rudman & Glick, 1999), and who we choose as leader (Castelli et al., 2009; Garfield et al., 2025; Todorov et al., 2005). Beyond partner selection, an accurate and precise representation of someone's competence enables predictions about which pieces of knowledge they possess. Such predictions are used in various aspects of our social life, such as allocating efforts in collaborative tasks (Xiang et al., 2023) or adapting an explanation to the audience's level during knowledge transmission (Gweon et al., 2014; Isaacs & Clark, 1987).

Assessing people's competence is a complex task: many potential cues correlate poorly with actual competence (e.g. non-verbal cues, Breil et al., 2020), and even more direct cues (e.g. solving a problem or giving a correct answer) are at best probabilistic. Someone can get lucky when guessing an answer or be unusually tired at test time (Jones, 1989). Moreover, when looking for someone to learn from, a learner begins with a knowledge gap, and inferring who is competent or knowledgeable when we are incompetent or ignorant ourselves is particularly difficult.

Like many aspects of social cognition, competence evaluation is an inferential process carried out under uncertainty, one whose success relies on following sound statistical principles (Griffiths et al., 2024; Quillien et al., 2023). Compared to the rich literature on beliefs and intentions attribution (e.g. Baker et al., 2017; Jara-Ettinger et al., 2016, 2020; Lucas et al., 2014; Quillien & Taylor-Davies, 2025), fewer studies have characterized how people infer competence, and how they select diagnostic evidence to update evaluation of competence, even if there is emerging evidence that people integrate information about others' competence in a rational, Bayesian fashion (Baker et al., 2017; Davis et al., 2025; Jara-Ettinger & Gweon, 2017; Xiang et al., 2026). Most of these studies, however, rely on artificial stimuli, such as descriptions of simulated agents who succeed or fail at different tasks (e.g. lifting boxes or solving mazes), and have artificial competence benchmarks, with performance being determined by the experimenters. While these paradigms allow for ex-

---

We acknowledge financial support from the Agence Nationale de la Recherche (ANR PACE ANR-25-CE28-0318 to Hugo Mercier, ANR-17-EURE-0017 to FrontCog, and ANR-10-IDEX-0001-02 to PSL). This work has also received support under the Major Research Program of PSL Research University "CultureLab" launched by PSL Research University and implemented by ANR with the references ANR-10-IDEX-0001.

The authors made the following contributions. Marius Mercier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; Olivier Morin: Conceptualization, Methodology, Writing – review & editing; Hugo Mercier: Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition; Tadeg Quillien: Conceptualization, Methodology, Writing – review & editing. † Olivier Morin, Hugo Mercier, and Tadeg Quillien contributed equally to this work.

Correspondence concerning this article should be addressed to Hugo Mercier. E-mail: hugo.mercier@gmail.com

perimental control, they often preclude participants from relying on a rich prior knowledge that characterizes real-world social judgements. Here, we examine competence evaluation in the domain of general knowledge, gathering data on participants' actual performance on general knowledge tests. This allows comparing participants' predictions with the actual knowledge of a population sample, and gives participants the opportunity to draw on priors regarding question difficulty and the distribution of ability in the population in the domain at hand.

Building on prior computational and empirical work on competence inference, we provide a systematic test of knowledge attribution in a naturalistic setting and compare people's behaviors with the predictions of a rational Bayesian model. We also investigate whether people's information-seeking strategies regarding others' competence are guided by the same rational principles.

1.1 Evaluation of competence

People rely on a wide range of cues when judging the competence of others, yet these cues differ considerably in their diagnostic value. For instance, facial cues are used to form first impressions of competence (Eisenbruch et al., 2024; Todorov et al., 2008; Todorov et al., 2015; Todorov & Oh, 2021), even though these impressions often correlate poorly with actual competence (e.g. Todorov et al., 2015). In a meta-analysis, Breil et al. (2020) showed that participants use several nonverbal cues to estimate the competence of others, including cues that are only weakly correlated (e.g. speech rate) or uncorrelated (e.g. eye contact) with competence.

What other cues, more correlated with actual competence, should a computational account of competence inference consider? Some cues are relevant only in specific contexts. For instance, age might not be a good indicator of competence in most contexts, but for young children, age is often correlated with competence–a cue that 4- and 5-year-olds rightly pick up on (Magid et al., 2018). Other cues that adults and children use more reliably correlate apply across contexts. Children as young as three use past accuracy in a task to predict future accuracy (Birch et al., 2008; Corriveau & Harris, 2009b; Harris et al., 2018; Koenig et al., 2004). Children and adults can base their judgements on the difficulty of the tasks completed (Dubourg et al., 2025) or attempted (Jara-Ettinger et al., 2015; Jara-Ettinger & Gweon, 2017). Toddlers and preschoolers are also more likely to trust someone who provided a good explanation (Castelain et al., 2018; Clegg et al., 2019). Adults likewise consider convincing explanations (Turpin et al., 2021) and the sharing of good ideas as signals of competence (Altay et al., 2020; Klopfenstein & Mercier, 2025). Children gradually show finer-grained inferences during their development, as children and adults alike judge individuals who act more efficiently as more competent (Kryven et al., 2021; Leonard et al., 2019; Török et al., 2023). Crucially, these cues are diagnostic: individuals who have been accurate tend to remain accurate in later, similar tasks (e.g. Himmelstein et al., 2021) owing to domain-specific knowledge or stable cognitive abilities (Breit et al., 2024; Kryven et al., 2021); the ability to produce high-quality explanations correlates with intelligence (Turpin et al., 2021) and knowledge (Lombrozo, 2006).

When cues of varying diagnosticity are available, people favor reliable cues. Young children put more weight on informants' past accuracy than on how familiar the informants are (Birch et al., 2008; Corriveau & Harris, 2009a), what the informants' accent is (Corriveau et al., 2013), their gender (Taylor, 2013), their hair color (Reyes-Jaquez & Echols, 2013), or whether they belong to the same group in a "minimal group" paradigm (Elashi & Mills, 2014). Adults' perception of a child's intelligence is influenced to a larger extent by the quality of the explanation the child provides than by the child's appearance (Blasi et al., 2015). Similarly, when judging credibility, a speaker's past accuracy is valued more than their expressed confidence (Tenney et al., 2007; Vullioud et al., 2017). Even though perceived competence from facial appearance seems to influence hiring decisions (Menegatti et al., 2021), participants consider these cues to be less useful and appropriate for hiring than other cues, such as education (Jaeger et al., 2022).

Among the diagnostic cues identified in the literature, past accuracy and task difficulty stand out as the primary dimensions for formal modeling (see e.g. Jansen et al., 2021; Jara-Ettinger & Gweon, 2017). Success on a task provides the most direct evidence of competence, while task difficulty modulates the inference to competence. By focusing on these two variables, we can isolate the fundamental logic of competence inference, while treating this as an abstraction from richer real-world judgments that integrate other cues, such as the ease with which a task is executed (Leonard et al., 2019).

1.2 Rational inference of competence

Inferring competence, even from reliable cues such as past success and task difficulty, is far from straightforward (e.g. Jones, 1989). Imagine you must choose between two informants based on their answer to a single question. The first correctly answers an easy question, while the second fails to answer a difficult one. Which informant would you choose? While the first informant was correct, their success on a simple question reveals very little about the depth of their expertise. Conversely, the second informant's failure is also hard to interpret; even an expert might fail to answer a particularly obscure question. How can people rationally infer the competence of others from such limited information?

In some contexts, judgments of competence rely on mentalizing, the ability to infer the mental states that drive others' actions. A prominent framework models this capacity as Bayesian inference: people have a generative model mapping the mental states of other agents to actions, and given what others do and perceive, observers can perform Bayesian inferences to infer beliefs and desires (Baker et al., 2011; Baker et al., 2017; Jara-Ettinger et al., 2020). Working within this framework, Aboody, Davis, et al. (2025) show that people can infer others' knowledge by treating choices as cost-benefit trade-offs: when the cost of different options varies as a function of how knowledgeable someone is, participants assume agents try to minimize costs and infer their knowledge from the options they choose (see also Jara-Ettinger et al., 2016; Jara-Ettinger & Gweon, 2017; Lucas et al., 2014; Xiang et al., 2023). Participants were able to infer knowledgeability from choices, even if they could not infer with precision what others knew (Aboody, Davis, et al., 2025). Here, we instead ask whether observers can infer knowledgeability from success or failure at a task, instead of from a decision.

We suggest that inferring competence from performance is a probabilistic inference problem, and we identify three key requirements for a rational inference of competence. First, observers must have a generative model linking competence to performance. Performance is inherently probabilistic: success in answering a question might be due to chance, and failure to a momentary lapse of attention (Jones, 1989; Rasch, 1960). Observers must thus treat performance as a cue to latent competence that can be reliable but is always imperfect. Second, when outcomes are noisy, rational observers should avoid putting excessive weight on a single observation and instead integrate it with their prior expectations about competence (e.g. Griffiths & Tenenbaum, 2006), although a single observation can still warrant a large belief update when it is highly diagnostic (e.g. success on an extremely difficult task). Third, observers should ideally integrate the noisy evidence with their prior expectations in a mathematically optimal manner, a process captured by Bayesian inference (Griffiths et al., 2024; Tenenbaum et al., 2006).

Close to our aim is a computational model by Shafto et al. (2012) which builds on results from the developmental literature showing that preschoolers integrate cues about an informant's past accuracy to decide who to learn from (e.g. Aboody, Lu, et al., 2025; Mascaro & Sperber, 2009; Pasquini et al., 2007). This is modeled as a joint Bayesian inference over the state of the world and over an informant's knowledge and helpfulness. Knowledge and helpfulness are treated as binary: knowledgeable informants are assumed to have the correct belief, whereas ignorant ones choose randomly among alternatives. Likewise, an agent is assumed either to be willing to disclose useful information (to be helpful) or not. This Bayesian framework captures children's trust judgments across diverse informant reliability conditions, showing how they simultaneously update their beliefs about the world and about informants' knowledge and intent. However, because it discretizes knowledge and does not model observed informant accuracy as a noisy function of graded competence, it does not support precise (e.g. item-level) knowledgeability estimations.

Another closely related model, recently introduced by Xiang et al. (2026), focuses on reputation management, but also integrates judgments of competence. They first introduce a generative model where a certain level of competence causes agents to pass or fail a test (which includes several questions), then use Bayesian inference to infer the competence level from the observed outcome. This model operates at the test level with a fixed difficulty (corresponding to the number of questions required to pass the test), by contrast with the current model which focuses on item-level answers.

Lastly, Jansen et al. (2021) model people's self-assessments as Bayesian inference regarding their own latent ability. This model combines prior beliefs with an item-level likelihood linking ability and difficulty, in order to understand the metacognitive mechanisms that track one's own performance. While their framework focuses on metacognition, the current work seeks to explain how people infer the competence of others, a process where metacognition is less central.

1.2.1 Computational Model. As suggested above, to infer another individual's competence, an agent should take into account the probabilistic nature of performance, should not overweight any single performance, and update their priors along Bayesian principles (assuming negligible processing costs, although see Todd & Gigerenzer, 2012). On the

basis of these requirements, we formulate a Bayesian model of competence inference that optimally integrates novel information with prior expectations. After observing whether an individual succeeds ($S = 1$) or fails ($S = 0$) at a task of difficulty β, the model updates its prior belief about the individual's competence θ. Defining the likelihood of success as $p(S \mid \beta, \theta)$, we can formalize the inferential problem that participants faced with the following Bayesian equation:

$$p(\theta \mid \beta, S) \propto p(S \mid \beta, \theta) \cdot p(\theta) \tag{1}$$

The likelihood $p(S \mid \beta, \theta)$ is given by a generative model that captures how competence and task difficulty jointly determine whether the agent succeeds. Since in our experiments the individuals are assigned to a task at random, we model task difficulty and competence as being independent, such that $p(\theta \mid \beta) = p(\theta)$, where $p(\theta)$ is the prior distribution over competence. When we observe someone succeeding at a task, the outcome can be interpreted as a combination of the individual being competent or the task being easy. In some ambiguous situations, observers must infer task difficulty and an agent's competence jointly from the observed outcome (Jansen et al., 2021; Jara-Ettinger & Gweon, 2017; Jones, 1989; Xiang et al., 2023). For the materials we use, however, participants have rich, and largely reliable (Dubourg et al., 2025), intuitions about the difficulty of trivia questions that are independent of whether a given person succeeds or fails (see also Supplementary Materials, Figure S5). Accordingly, our model treats difficulty as known and uses success or failure to infer the individual's competence. We therefore do not model how observers infer difficulty which would likely be highly domain-specific. By contrast, the mapping from difficulty to competence in our Bayesian framework is arguably more general: it can be applied across domains independently of the particular mechanism by which difficulty judgments are formed.

First, we assume that $p(\theta)$ is given by a normal distribution with mean μ and variance σ^2: $\theta \sim \mathcal{N}(\mu, \sigma^2)$. Second, we use the following generative model in which an individual is more likely to succeed if their competence θ is above the task difficulty β. Nonetheless, outcomes are stochastic: even individuals whose competence is slightly below the task difficulty may succeed by chance. Formally, a commonly used model to estimate test-takers' latent ability is the one-parameter Item Response Theory model (or Rasch model, Rasch, 1960). We define the likelihood function $p(S \mid \beta, \theta)$ as a normal-ogive IRT model, which corresponds to the cumulative distribution function (Φ) of a normal distribution centered around β with a standard deviation ε:

$$p(S \mid \beta, \theta) = \big[\Phi((\theta - \beta)/\varepsilon)\big]^{S} \big[1 - \Phi((\theta - \beta)/\varepsilon)\big]^{1-S} \tag{2}$$

This function computes the probability of an individual succeeding at a task of difficulty β given a competence level θ. The parameter ε is a free noise parameter (i.e., randomness in task success) that controls how sharply success increases with $\theta - \beta$: smaller ε yields more deterministic outcomes. More intuitively, the more an individual's competence is above the task's difficulty, the higher the probability that the individual successfully completes the task. By definition, when $\theta = \beta$, the probability of success is at chance

($\Phi(0) = 0.5$). Our model is similar to that of Jansen et al. (2021) in that we both adopt a Rasch model likelihood, which provides a validated psychometric framework to link competence to item difficulty. Figure 1 illustrates the several parts of the models along with the corresponding task for the participants.

The proposed computational model addresses the three requirements for rational competence inference. First, we define a generative model that links performance and competence through our likelihood function (Equation 2). This assumes that performance is probabilistically influenced by the latent competence, with the noise parameter ε capturing the inherent uncertainty in this link. Second, the model incorporates a prior distribution $p(\theta)$ that represents observers' baseline expectations about competence levels. Third, the model implements Bayesian inference, which provides statistically optimal belief updating given the specified priors and likelihood function, thereby adhering to the principles of the ideal observer model (Geisler, 2011). Together, these components yield precise predictions about how rational observers should update their competence judgments across different patterns of success and failure.

1.2.2 Overview of empirical tests. The present research aims to formalize the underlying cognitive mechanisms of competence inference and test the robustness of this ability in more complex scenarios. We do so in three studies that each provide a unique contribution. We focus on inference of knowledgeability, a facet of competence, which allows for reliable assessment through trivia quizzes. Recent research has demonstrated that people can accurately infer overall knowledge of other participants after observing them solve a single trivia question (Dubourg et al., 2025). This paradigm provides an interesting test of our model, as Bayesian models are most diagnostic under information-limited conditions, where the balance of prior and likelihood can be cleanly observed (Baker et al., 2009; Vul et al., 2014). Study 1 provides an initial test of the model's plausibility by fitting it to the original Dubourg et al. (2025) dataset and demonstrates that it accounts for participant judgments more effectively than plausible heuristic alternatives. Second, we test the generality and robustness of our model. While Dubourg et al. (2025) only presented participants with correct answers, real-world inferences often rely on observation of failure. Study 2, therefore, presents a pre-registered extension in which individuals can also fail at a task, allowing us to test whether participants could accurately predict the competence of others from incorrect answers. Third, we shift from passive judgement to active information seeking. To more closely mirror how people assess the competence of others in the real world, Study 3 introduces a new Information Search paradigm. Here, we ask whether participants can seek the most diagnostic evidence to update their competence evaluations.

2 Study 1

This study re-analyses the open dataset by Dubourg et al. (2025) to test a computational model of competence inference. We hypothesize that participants' behavior is best explained by a Bayesian model and compare it with lesioned models corresponding to different heuristics. For this study, the computational models were not pre-registered. How-

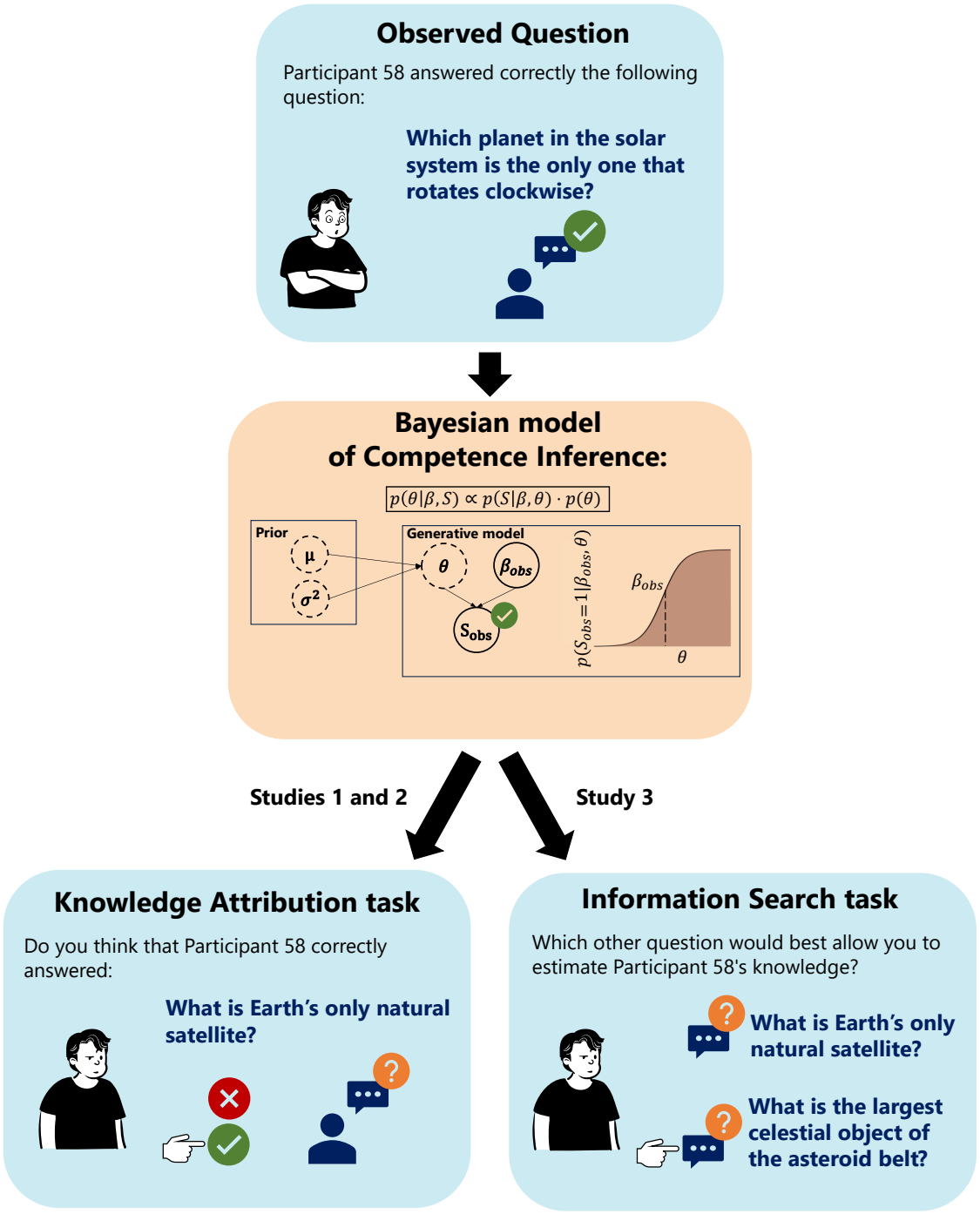

Figure 1. Illustration of the different tasks and the Bayesian model of competence inference. We represent here the Bayesian model which infers the competence θ of an individual after observing them correctly or incorrectly answering a question (S_{obs}) of a given difficulty β_{obs}. The dashed circle means that the value is not observed by the model. The model has a prior over competence which follows a normal distribution with a mean μ and a variance σ^2. We also illustrate how inferred competence is linked with the probability of success for a given difficulty. Note that when inferred competence is exactly equal to the difficulty, the probability of success is 50%; it increases when θ is superior to β and decreases when θ is inferior. The competence θ is inferred via Bayes' rule (Equation 1) by inverting the generative model. The models and procedures are further detailed in the Methods of the corresponding studies.

ever, all code and analyses for the three studies are openly available[1]. The methods from Dubourg et al. (2025) are described here in depth as they help understand the behaviors the present models will seek to emulate, and because they are used (with some modifications) to collect new data in Study 2.

2.1 Methods

2.1.1 Participants. Dubourg et al. (2025) recruited 931 U.S participants via the online platform Prolific. They excluded participants who failed the attention check (N = 11) and participants who gave an answer in contradiction with the instruction (e.g. saying that someone can't answer a question although it is indicated that they answered it, N=72). 848 participants were thus included in the analysis (451 women, 422 men, M_{age} = 43.8, SD_{age} = 13.12).

2.1.2 Behavioral task. The behavioral task consisted of two consecutive phases: a competence evaluation phase and a questionnaire. In the first phase, each participant was randomly assigned one quiz theme–American History, Superheroes, or Solar System– each comprising 15 questions (see Materials below). For each of five virtual individuals (simulated participants), Dubourg et al. (2025) revealed that the individual had answered one randomly selected item correctly (e.g "How old is the Sun?"). Participants saw the text of that item but not its answer. We call this item, which is the only piece of information that they have about the participant, the observed question.

After having seen the observed question, participants predicted for each of the remaining 14 questions in the theme whether that individual would answer it correctly (binary scale: "Yes"/"No"). We call each of these items a "new question." Note that our analyses are conducted at the level of observed-new question pairs (i.e., for each observed question, the 14 predictions on the new questions). Participants were further asked to rate the difficulty of the observed question between 0 and 100: "Out of 100 participants, how many do you think got the question right?" (reverse-coded as 100 - rating). This procedure was repeated five times per participant, each time with a new randomly chosen question and a new individual label (Participant B, Participant C, etc.), making it explicit that different individuals were being evaluated.

In the second phase, participants completed the full 15-item quiz for the theme they had just evaluated. The answers were open-text boxes and participants were instructed that they could say that they did not know the answer.

2.1.3 Materials. Three different general-knowledge quizzes were used (Solar System, American history and Superheroes), each comprising 15 trivia questions of varying difficulty. Examples of questions include: "What is the name of the first man that stepped on the Moon?" and "What is the largest celestial object of the asteroid belt?" (Solar System); "What is the year of the beginning of the American Civil War?" and "Which President was in office between 1893 and 1897?" (American History); and "Who killed Superman in the 1993 comic?" and "Who created the character of Iron Man?" (Superheroes). Full materials are available in the Supplementary Materials (SM, section 1).

[1] https://osf.io/aecyr/files/github?view_only=2b8fc0d1c99e450f8b8534e2aab83a92

2.1.4 Computational Models. ***Input data for the models.*** Each question is assigned a difficulty score β. We first determine each question's judged difficulty by averaging participants' difficulty ratings and dividing the ratings by 100. We then applied a logit (qlogis) transformation [2] to map these scores from (0,1) to an unbounded scale.

Main Bayesian model. Our computational framework assumes that, in order to do the task, participants first infer the competence θ of an individual from the difficulty of the observed question β_{obs} and their success S_{obs} by using the previously described Bayesian inference $p(\theta \mid \beta_{obs}, S_{obs})$. Second, participants have to predict the probability of success of the individual on a new question of difficulty β_{new}. Using the law of total probability, this can be computed by weighting the probability of success given each possible competence level by the posterior probability of that competence level derived from the previous observation:

$$P(S = 1 \mid \beta_{new}, \beta_{obs}, S_{obs}) = \int \Phi((\theta - \beta_{new})/\varepsilon) p(\theta \mid \beta_{obs}, S_{obs}) d\theta \tag{3}$$

The Bayesian model has four free parameters that we fit to the data: the mean μ and variance σ^2 of the prior $p(\theta)$ over competence, the noise ε, and the temperature τ of the softmax function that turns probability estimates into binary answers (see below).

We compare our main model with two alternative heuristic models: a "Threshold" heuristic and an "Anchor" heuristic.

Threshold heuristic model. The Threshold heuristic formalizes a simple rule: if the individual succeeded, their competence is at least the item's difficulty; if the individual failed, their ability is no higher than that difficulty, and all values of competence are otherwise considered equally plausible. It is a lesioned version of the Bayesian model which has the following two constraints. First, Competence follows a uniform prior on [-4,4] [3]. Second, performance is assumed to be deterministic ($\varepsilon = 0$): the individual answers correctly if and only if $\theta \geq \beta$. It follows that observing a correct response implies $\theta \geq \beta_{obs}$; the posterior is uniform on $[\beta_{obs}, 4]$ and zero elsewhere. An incorrect response implies $\theta < \beta_{obs}$; the posterior is uniform on $[-4, \beta_{\text{obs}}]$ and zero elsewhere. Under the Threshold heuristic, the predicted success probability on a new item equals the posterior mass of competence above that item's difficulty:

$$P_T(S_{\text{new}} = 1 \mid \beta_{\text{new}}, \beta_{\text{obs}}, S_{\text{obs}}) = \int_{\beta_{\text{new}}}^{4} p_T(\theta \mid \beta_{\text{obs}}, S_{\text{obs}})\, d\theta \tag{4}$$

which is obtained from the general predictive formula by substituting the heuristic posterior $p_T(\theta \mid \beta_{\text{obs}}, S_{\text{obs}})$ and setting $\varepsilon = 0$, thereby turning the probit link $\Phi\big((\theta - \beta_{\text{new}})/\varepsilon\big)$ into the indicator $\mathbf{1}\{\theta \geq \beta_{\text{new}}\}$. The threshold heuristic model has one free parameter: the temperature τ of the softmax function.

[2] This transformation is inspired by the way difficulty is operationalized in frameworks like Item Response Theory (Rasch, 1960). Intuitively, difficulty is not a bounded scale because even two questions that are scored as having 100% difficulty (everyone in a sample of respondents fails to answer) might still not have the same true difficulty. The qlogis transform is given by: $qlogis(p) = log(p/(1-p))$.

[3] Theoretically, the values of θ can go from $-\infty$ to ∞ but we reduce the interval from -4 to 4 for computational reasons.

Anchor heuristic model. Whereas the Threshold heuristic retains a full posterior over competence, the Anchor heuristic collapses inference to a single-point estimate $\hat{\theta}$. After observing an item of difficulty β_{obs} and its outcome S_{obs}, the point-mass distribution is given by: $\hat{\theta} = [\beta_{obs} + \Delta]^{S}[\beta_{obs} - \Delta]^{1-S}$ where Δ is a free parameter above 0. Performance is again deterministic ($\varepsilon = 0$), so the predicted success probability on a new item of difficulty β_{new} is simply:

$$P_A(S_{\text{new}} = 1 \mid \beta_{\text{new}}, \beta_{\text{obs}}, S_{\text{obs}}) = \mathbf{1}\{\beta_{\text{new}} \leq \hat{\theta}\} \tag{5}$$

Intuitively, the model assumes that participants "anchor" their competence belief a small distance Δ away from the difficulty of the observed question — Δ above the difficulty if the question was answered correctly, or Δ below the difficulty if the question was answered incorrectly. The model then predicts deterministic success for items no harder than that anchor and failure otherwise. The Anchor heuristic model has two free parameters: Δ and τ.

Modelling Participants' Binary Answers. All models estimate the probability that an individual will answer a given question correctly. However, participants were asked to provide binary answers. We assume that participants generate these answers stochastically on the basis of internal probability estimates. We model this stochastic judgment process using a softmax choice rule applied to a Bernoulli decision. The probability that a participant answers "Yes" to a question in the evaluation phase is defined as:

$$P(\text{"Yes"} \mid \beta_{\text{new}}) = \frac{\exp(p_{model}/\tau)}{\exp(p_{model}/\tau) + \exp((1 - p_{model})/\tau)} \tag{6}$$

Simplifying this equation, we obtain:

$$P(\text{"Yes"} \mid \beta_{\text{new}}) = \frac{1}{1 + \exp((1 - 2p_{model})/\tau)} \tag{7}$$

where $p_{model} = P(S = 1 \mid \beta_{\text{new}})$ for a given model. The temperature parameter τ is a free parameter that controls decision randomness: lower values of τ result in more deterministic (less random) choices, while higher values lead to increased randomness.

2.2 Results

All statistical analyses were conducted in R (v.4.5.2), using RStudio (2026.01.0+392). We first describe participants' performance on the task. Second, we fit the models, optimizing the free parameters on all data, and compare them to establish which model better accounts for participants' behavior. Lastly, we present various exploratory analyses, including recovering the best parameters for each participant and conducting robustness checks. All regression models included a fixed effect for the intercept by default; we do not report intercept estimates because they are not informative for our hypotheses. They can be easily retrieved from our open code and data.

2.2.1 Behavioral Results. Dubourg et al. (2025) report that participants were able to infer the knowledge of individuals from a single piece of information. Participants' prediction of the score of an individual correlated with the individual's actual score on the quiz. Aggregating across participants, the average estimated question difficulty was highly correlated with the true difficulty of the question (r = 0.95; Pearson's product-moment correlation t(43) = 19.79, p < .001), see Figure 2A. Replicating Dubourg et al. (2025), we modeled predicted difficulty as a function of true difficulty at the trial level with random intercepts for questions and random intercepts and slopes by participants. Participants' predictions tracked the true difficulty ($\beta = 0.49$, t(58.84) = 19.14, p < .001). The authors highlight that knowing the answer to a question was not necessary to estimate its difficulty, as even the lowest-performing 30% of the participants (who were not provided with answers for unsolved questions) were able to accurately guess the true difficulty of a question.

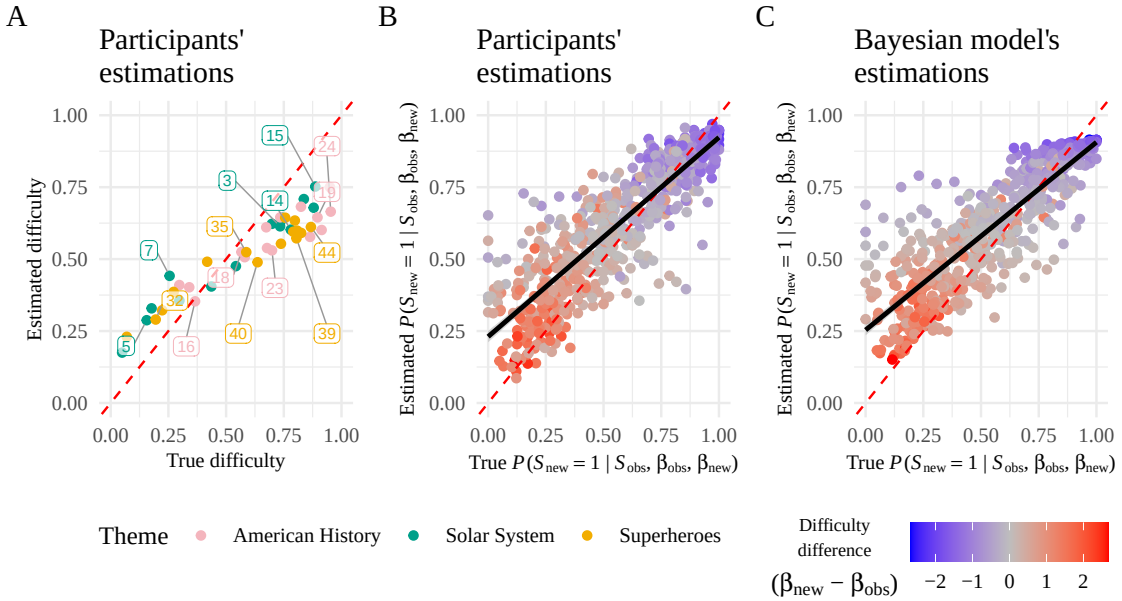

Figure 2. (**A**) Each question's estimated difficulty as a function of its true difficulty. Numbers refer to the question ID (see SM, section 1). The red dashed line represents a perfectly accurate estimation. (**B**) Participants' estimated probability of answering a new question correctly, conditioned on the performance on the observed question, as a function of true conditional probabilities. (**C**) Our Bayesian model's estimated conditional probabilities as a function of true conditional probabilities. For (**B**) and (**C**), each data point corresponds to the average values for a pair of two questions: one for which the individual's performance is observed, and one for which the individual's performance has to be guessed. The red line represents a perfectly accurate prediction, while the dark line represents the fit of the linear model.

In a re-analysis of the data, we found that participants can predict which piece of knowledge an individual possesses from the information provided. We used the data collected during the questionnaire phase to compute the conditional probability of correctly

answering a new question given that another question had been successfully answered. This allowed us to compare participants' average prediction on the competence evaluation task with an objective benchmark (see Figure 2B). We fitted a linear model with the true conditional probability as the independent variable and the average estimated conditional probability as the dependent variable. Participants were able to predict the true conditional probability with great precision ($\beta = 0.87$, t(628) = 43.78, p < .001, Adjusted $R^2 = 0.75$). At the trial level, participants' answers were binary predictions of the individual answering a question. To ensure that the effect was not merely a wisdom-of-the-crowd artifact, we ran a logistic mixed-effects model with binary predictions as the dependent variable and true conditional probability as the independent variable with random intercepts for the observed question and the evaluated question, plus random intercepts and slopes by participants. This model indicated that participants' predictions of success increased with the true conditional probability (b = 5.53, z = 40.86, p < .001). This result shows that participants can make accurate conditional probability judgments. In the next section, we ask whether this good performance is best explained in terms of normative Bayesian inference or can be explained by simpler heuristics.

2.2.2 Modelling Results. For each model, we estimated the best-fitting free parameters by maximizing the log likelihood of the participants' answers, given the models and sets of parameters. All models were given the information that the virtual individuals succeeded ($S_{\text{obs}} = 1$) in answering a question of a difficulty β_{obs}, and had to predict the success of this individual on a new evaluated question of difficult β_{new}. Based on the maximum log likelihood for each optimized model, we calculated the Bayesian information criterion (BIC) which penalizes more complex models. Our main model, fitted to all the data, obtained a BIC of 67170, better than both alternative models ($BIC_{Threshold} = 76192$; $BIC_{Anchor} = 74379$; lower BIC values indicate better fit). The difference between our Main Model and the second best model is 7209, which indicates very strong evidence in favor of our Main Model per standard interpretation criteria (Raftery, 1995). Table 1 indicates the best-fitting parameters for each model.

Figure 3 represents participants' average prediction for each pair of questions (630 observed–new question pairs) compared to the three models' predictions. For each computational model, we fitted a linear regression with the model's predicted probability as the independent variable and participants' average prediction as the dependent variable on all pairs. Our main model explains 90% of the variance in participants' prediction ($\beta = 0.95$, t(628) = 73.48, p < .001, Adjusted $R^2 = 0.90$). The Threshold model explains 59% of the variance ($\beta = 0.77$, t(628) = 30.21, p < .001, Adjusted $R^2 = 0.59$) and the Anchor model explains 39% of the variance ($\beta = 0.63$, t(628) = 20.11, p < .001, Adjusted $R^2 = 0.39$). The main model was more strongly correlated with participants' answers (r = 0.95) compared to the Threshold model (r = 0.77) and the Anchor model (r = 0.63).

2.2.3 Exploratory and robustness analyses. The "Threshold heuristic" differed from the main model in two ways: ε was set to 0 for a deterministic cut-off and the prior was set to a uniform distribution. To establish the relative contribution of having a probabilistic likelihood and a normal prior, we conducted an ablation analysis. When modifying the Bayesian model by setting ε to 0 (and keeping a normal prior) we found a BIC of 69672, and when setting the prior to a uniform distribution (and letting ε vary as

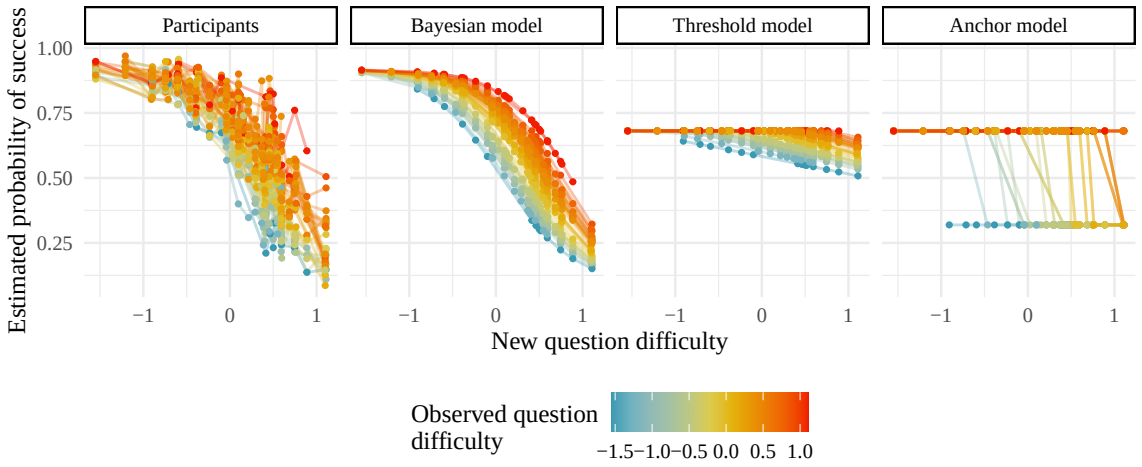

Figure 3. The participants panel shows participants' estimation of the probability of succeeding on a new question depending on its difficulty, given that a previous question (the observed question) has been answered successfully. The predictions made from the same information are grouped together by a line and colored as a function of the difficulty of the observed question. Each point represents an observed–new question pair. The three other panels represent the predictions made by the Bayesian model and the two heuristic models.

Table 1
Model comparison results for Study 1

Model	LLmax	BIC	μ	σ^2	ε	Δ	τ
Bayes	-33563	67170	0.01	0.71	0.66		0.42
Threshold	-38091	76192					1.32
Anchor	-37178	74379				0.58	1.32

a free parameter) we found a BIC of 71278. This indicates a better fit to participants' behavior than the Threshold heuristic ($\Delta BIC_{\text{vs threshold}}$ = -6520 and -4914), but still worse than the full Bayesian model ($\Delta BIC_{\text{vs Bayes}}$ = 2503 and 4108), thus supporting a distinct contribution of both components.

To confirm that the results of our confirmatory analyses do not hinge on the fitting methods or on how item difficulty is represented, we conducted additional analyses (see Supplementary Materials for a detailed report).

First, we re-fitted the models in Stan (Carpenter et al., 2017) and compared them via Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO; Vehtari et al. (2017); see SM, section 2). The Bayesian model showed a credible higher out-of-sample predictive accuracy than the Threshold heuristic ($\Delta elpd$ = -4,524.85, SE = 89.21); the Anchor heuristic was not re-fit in Stan because its deterministic cut-off induces a non-differentiable likelihood.

Second, we performed an individual-based analysis where we fitted each candidate model for each participant, allowing them to have different parameters (see SM, section 3).

We computed a BIC for each participant and for each fitted model, we report the average BIC along with the standard deviation for each model. The main model fit was significantly better than the Threshold model ($BIC_{\text{Bayes}} = 78.90 \pm 21.12$; $BIC_{\text{Threshold}} = 87.77 \pm 16.68$; t(895) = -15.45, p < .001, paired t-test) and the Anchor model ($BIC_{\text{Anchor}} = 83.19 \pm 17.83$; t(895) = -10.32, p < .001, paired t-test).

Lastly, we ran robustness checks that replaced judged difficulty with an objective estimate of difficulty (see SM, section 4). Our main analyses use mean perceived difficulty because we focus on how people make inferences given an estimate of item difficulty; how those estimates are formed is beyond the scope of this paper. To ensure this choice does not drive our results, we re-ran all analyses using item difficulties estimated from a Rasch model (Rasch, 1960) fit to participants' responses. The Bayesian model remained the strongest predictor of participants' predictions, outperforming the Threshold and Anchor heuristics.

The results show that participants can accurately estimate not only the overall knowledge level of the virtual individual but also the probability of correctly answering each question given the observed question. Our modelling results show that participants likely use a generative model in which the estimated competence affects the probability of solving a problem following Bayesian principles.

3 Study 2

The previous study has shown that participants' inferences of the knowledgeability of others can be accurately modeled as a rational Bayesian process. One limitation of the previous study is that it only included observations in which the individual successfully answered a question. Study 2 is thus a replication and extension of the previous study: in addition to replicating the paradigm from Study 1, we investigate the inferences people make about individuals who fail to answer a question. We predicted that participants' behavior would still be best captured by the optimal Bayesian Model. More precisely, we put forward the following hypotheses:

H1. The predictions of the main Bayesian model are positively correlated with participants' average predictions.

H2. When fitted to the aggregated dataset, the main Bayesian model better captures participants' behavior than the two alternative heuristic models.

The procedures, data collection, analysis plan and the models (including implementation code) were pre-registered[4]. All code and data are available[5].

3.1 Methods

3.1.1 Participants. 899 U.S participants were recruited via the online platform Prolific. Two exclusion criteria were pre-registered. First, participants who failed the attention check were excluded (N = 19). Second, we excluded participants who gave an

[4]https://osf.io/aecyr/files/osfstorage?view_only=2b8fc0d1c99e450f8b8534e2aab83a92
[5]https://osf.io/aecyr/files/github?view_only=2b8fc0d1c99e450f8b8534e2aab83a92

answer that contradicted the instructions (e.g. saying that someone had answered a question successfully although it was indicated that they had failed to answer it, N=108). 772 participants were thus included in the analysis (403 women, 367 men, 2 preferred not to say, $M_{age} = 41$, $SD_{age} = 12.65$).

3.1.2 Behavioral Task. This study followed the design of the evaluation phase of Study 1 (see section 2.1.2). Participants were not presented with the questionnaire phase. The materials used were the same as in Study 1 (see SM, section 1). As with Study 1, participants were asked to evaluate five virtual individuals after being presented with a single piece of information about them: the answer to a single question (all questions were drawn from the questions used in Study 1, which fell into three themes). The only difference with Study 1 was a within-subject manipulation: on each trial, participants were randomly told that the individual either failed to answer the observed question or successfully answered it. Participants were then asked to evaluate whether the individual got the other 14 questions of the same theme right or wrong and to evaluate the difficulty of each observed question following Study 1 procedures.

3.1.3 Computational Models. The main Bayesian model and heuristics used in Study 1 were pre-registered and re-used in this study.

3.2 Results

3.2.1 Behavioral Results. Participants' average estimation of the mean difficulty of a question was highly correlated with the true difficulty of the question as estimated from Study 1 (r = 0.95). Participants' judgements of difficulty in Study 2 were almost identical to those in Study 1 (r = 0.98). At the trial level, a linear mixed-effects model with a random intercept for the observed question and by-participant random intercepts and slopes showed that predicted difficulty significantly tracked true difficulty ($\beta = 0.47$, t(64.82) = 18.28, p < .001).

We computed the true conditional probability of answering the evaluated question given the individual's answer to the observed question, using participants' answers to trivia questions from the second phase of Study 1. We then fitted a linear model with the true conditional probability as the independent variable and participants' average judgement for each observed-new question pair as the dependent variable (N = 1,260: 630 pairs × 2 observed outcomes; see Figure 4). Results indicate that participants are able to predict true conditional probability with very high precision ($\beta = 0.91$, t(1258) = 77.13, p < .001, Adjusted $R^2 = 0.83$). Performance was a little bit higher when restricting our analysis to trials with only information about a failure ($\beta = 0.92$, t(628) = 57.63, p < .001, Adjusted $R^2 = 0.84$), compared to trials with only information about a success ($\beta = 0.86$, t(628) = 41.71, p < .001, Adjusted $R^2 = 0.73$). We also ran a trial-level mixed-effect model with participants' binary predictions as the dependent variable. A logistic model with random intercepts for the observed question, the evaluated (new) question, the observed outcome, and by-participant random intercepts and slopes, further showed that participants accurately predicted the conditional probability of answering a new question given the outcome on the observed question (b = 2.16, z = 13.16, p = < .001).

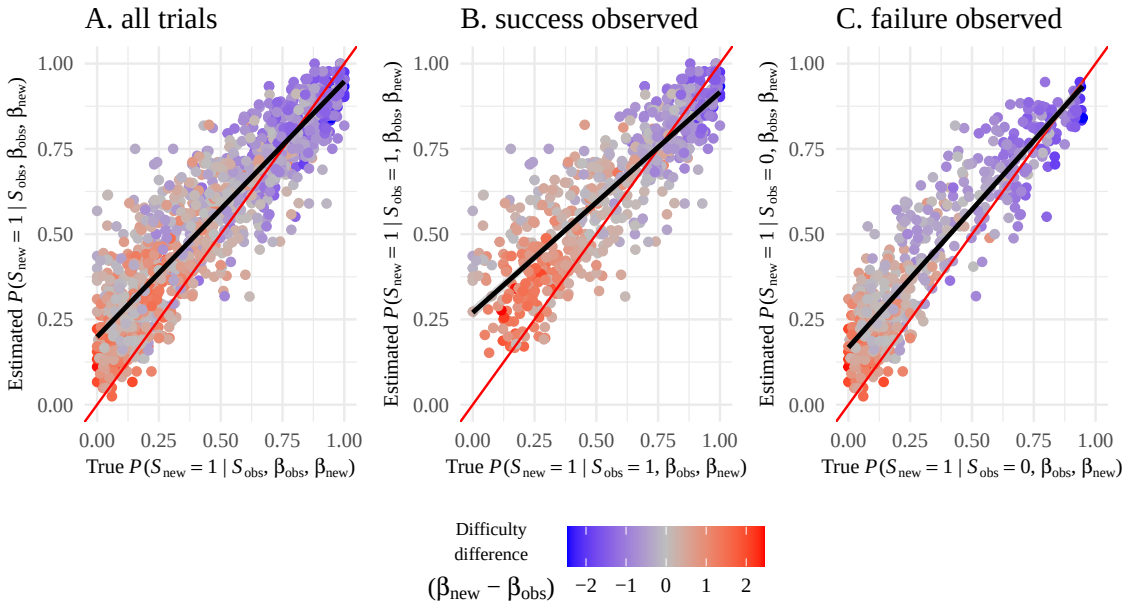

Figure 4. (**A**) Participants' estimated probability of answering a new question, given performance on a previous question (the observed question), as a function of the true conditional probability, including all trials. (**B**) Trials in which the observed question is correctly answered. (**C**) Trials where the observed question is incorrectly answered. Each data point is an observed-new question pair. The red line represents the optimal prediction while the dark line represents the fit of the linear model.

3.2.2 Modelling results - confirmatory analyses. As in Study 1, we used maximum likelihood estimation to find the best-fitting parameters for each model (see Table 2). We compared the three models' predictions with the average judgement of participants by fitting three linear models (see Figure 5). Supporting H1, our main model was strongly correlated with participants' behavior ($\beta = 0.94$, t(1258) = 101.22, p < .001). In support of H2, our main model obtained a BIC of 61892 indicating a better fit to the data compared to the two heuristic models ($BIC_{\text{Threshold}} = 69950$, $BIC_{\text{Anchor}} = 67220$). We also found that the main model explained 89% of the item-level variance (Adjusted $R^2 = 0.89$), whereas the Threshold model explains 36% of the variance ($\beta = 0.60$, t(1258) = 26.47, p < .001, Adjusted $R^2 = 0.36$), and the Anchor model explains 56% of the variance ($\beta = 0.75$, t(1258) = 40.15, p < .001, Adjusted $R^2 = 0.56$).

Table 2
Summary of the fit results for Study 2.

Model	LLmax	BIC	μ	σ^2	ε	Δ	τ
Bayes	-30924	61892	0.32	0.59	0.69		0.45
Threshold	-34969	69950					1.41
Anchor	-33599	67220				0.54	1.27

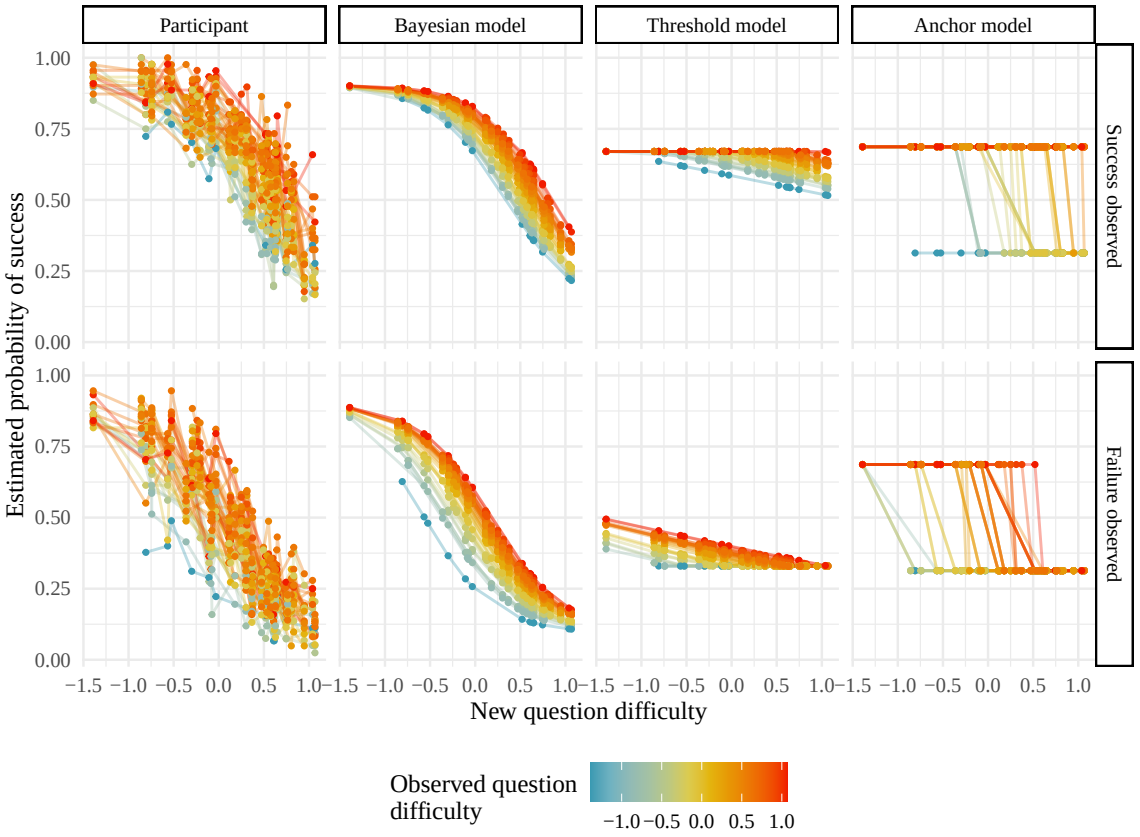

Figure 5. The participants panel shows participants' estimation of the probability of succeeding on a new question depending on its difficulty, given that another question (the observed question) has previously been answered successfully (upper panel) or unsuccessfully (lower panel). The predictions made from the same information are grouped together by a line and colored in function of the difficulty of the observed question. Each point represents an observed-new question pair. The top row includes only the trials when success was observed and the bottom row includes only the trials when failure was observed. The three other panels represent the predictions made by the Bayesian model and the two heuristic models.

3.2.3 Exploratory and robustness analyses. We report below a summary of the exploratory and robustness analyses performed, see Supplementary Materials for details. We conducted the same analyses as in Study 1.

We replicated the ablation analysis and found that removing either the probabilistic likelihood (BIC = 63955, $\Delta BIC_{\text{vs Bayes}}$ = 2063) or the normal prior (BIC = 66589, $\Delta BIC_{\text{vs Bayes}}$ = 4697) led to a worse fit compared to the full Bayesian model.

Using PSIS-LOO to compare our models, we found that the Bayesian model had a higher out-of-sample predictive accuracy than the Threshold model ($\Delta elpd$ = -4,040.93, SE = 84.66; see SM, section 2).

We also estimated the best free parameters for each participant (see SM, section 3). Our main model explained the behavior of participants significantly better than the Threshold heuristic (BIC_{Bayes} = 76.85 ± 21.46; $BIC_{\text{Threshold}}$ = 88.98 ± 15.37; t(771) = -19.65, p < .001, paired t-test) and Anchor heuristic (BIC_{Anchor} = 84.75 ± 17.61; t(771) = -16.11, p < .001, paired t-test).

Lastly, we replicated our main analyses by providing models with true difficulty and found that the Bayesian model remained the best-fitting model, outperforming the Threshold and Anchor heuristics (see SM, section 4).

This study replicates the findings of Study 1 and extends them to cases in which the observed question was not successfully answered.

4 Study 3

The previous studies have shown that participants are able to rationally infer the competence of an individual based on a single piece of information, using this estimation of competence to accurately attribute knowledge to others. Bayesian updating is most clearly demonstrated when participants must draw inferences from limited data.

Study 3 uses a different paradigm, one that asks participants to query information that would help them better assess competence. This paradigm provides an important test of Bayesian updating in more naturalistic conditions where people are not limited to a single piece of information. The Bayesian framework allows one to quantify the information value of a query, allowing us to derive principled normative predictions about participants' choices.

Assuming that people are interested in gathering more information about others' competence, are they able to compute the expected amount of information gained if they observe the individual answering a new question, and as a result select the most informative question to ask?

Assessing the rationality of people's information search is typically done by comparing people's choices to a normative model (Dubey & Griffiths, 2020; Klayman & Ha, 1987; Nelson, 2005; Oaksford & Chater, 1994; Tsividis et al., 2014). We adopt an Optimal Experimental Design (OED) framework (Liefgreen et al., 2020; Nelson, 2005; Quillien, 2023). OED models are composed of two main parts: a specification of the inferences an

individual should make given an observation (identified by Studies 1 and 2), and a measure of the amount of information gained following such an observation.

In this study, we present participants with one piece of information about an individual's competence. We then present participants with a discrete choice task in which they have to choose which new question to ask that individual. The Bayesian model is used to formalize participants' estimation of competence if they asked each question, which allows us to design a normative model to quantify the Expected Information Value of each query. We pre-registered the following hypotheses:

H1. The participants' odds of selecting a query are significantly correlated with the Expected Information Value of the query as computed by the normative model.

H2. Participants' information search behavior depends on the information they have been provided about the previous performance of the individual they are evaluating.

The procedures, data collection, analysis plan and the models (including implementation code) were pre-registered[6]. All code and data are available[7].

4.1 Methods

4.1.1 Participants. Three hundred and one U.S participants were recruited via the online platform Prolific. Five participants who failed the attention check were excluded for a final sample of 296 participants (155 women, 140 men, $M_{age} = 39.2$, $SD_{age} = 13.4$).

4.1.2 Behavioral task. The first part of the task was similar to the previous studies. Participants were introduced to a virtual individual who had answered, rightly or wrongly, a question on one of three themes. For the observed questions, we used a subset of five questions from the materials of the previous studies (see SM, section 1 for Materials). Then, in contrast with the previous studies, participants were told that they had to select an additional question in order to gather more information to assess that individual's level of knowledge or ignorance. Participants were told to imagine that we would give them feedback about whether the individual got the selected question right or wrong. Note that we were interested in which question they would select, so no feedback was provided. They could select one question among a set of six pre-selected questions (one set per theme, see SM, section 1). We refer to these selectable questions as "queries." This procedure was repeated 10 times: participants were randomly assigned to a theme and were presented with all ten virtual individuals for this theme (each of the five observed questions answered correctly or incorrectly). Figure 1 provides an illustration of the task.

4.1.3 Computational Models. After observing one performance of the virtual individual, participants had to select which question they would like to reveal in order to improve their estimation of competence. We formalize a normative model with two components: (a) an ideal observer which specifies the inference made after an observation and (b) an ideal search model that scores each candidate query by its Expected Information Value (assuming that inferences is made in the way specified by the ideal observer model).

[6] https://osf.io/aecyr/files/osfstorage?view_only=2b8fc0d1c99e450f8b8534e2aab83a92

[7] https://osf.io/aecyr/files/github?view_only=2b8fc0d1c99e450f8b8534e2aab83a92

Ideal observer model. The ideal observer model uses the difficulty of the observed question (β_{obs}) and the observed success or failure (S_{obs}) to update its estimate of competence $p(\theta \mid \beta_{\text{obs}}, S_{\text{obs}})$. The model is strictly identical to the Bayesian model presented in the previous studies that provide evidence that this model is a good approximation of people's inference. For this study, we do not re-fit the free parameters; instead we use the fixed parameters established in our pre-registration ($\mu = 0.15$, $\sigma^2 = 0.65$ and $\varepsilon = 0.7$), which were derived from the best-fitting values of Studies 1 and 2. Thus, before seeing any information, the model has a prior over competence of $\theta \sim \mathcal{N}(\mu = 0.15, \sigma^2 = 0.65)$.

Ideal search model. The ideal search model aims to quantify the information gained from selecting a query. We assume that once they see new information (i.e observing a success or failure S_{new} on a new question of difficulty β_{new}) they will perform a Bayesian update as described in the previous section.

We measure the information gained by observing a new question answered by the virtual individual as the Kullback-Leibler divergence (D_{KL}; Kullback & Leibler, 1951) of the previous estimation of competence $p(\theta)$ from the new posterior distribution $p(\theta')$, given by:

$$D_{\text{KL}}(p(\theta') \mid\mid p(\theta)) = \int p(\theta') \log\left(\frac{p(\theta')}{p(\theta)}\right) d\theta' \tag{8}$$

When selecting a query, participants do not yet know how their competence estimation will be affected. However, they can calculate the Expected Information Value (EIV) of each potential query. They do so by simulating two scenarios—one in which the individual answers the new question correctly ($S_{new} = 1$), and one in which they answer incorrectly ($S_{new} = 0$)—and then weighting these outcomes by their respective probabilities (given their estimation of the difficulty β_{new} of the question). This procedure can be formally expressed as:

$$EIV(Query) = \sum_{s \in \{0,1\}} D_{\text{KL}}\left(p(\theta' \mid \beta_{\text{new}}, S_{\text{new}} = s) \mid\mid p(\theta)\right) P(S_{\text{new}} = s \mid \beta_{\text{new}}, \theta) \tag{9}$$

In summary, the normative model updates its estimation of competence once with the initial observation (β_{obs}, S_{obs}). Then, for each candidate query with difficulty β_{new}, it computes the Expected Information Value (Equation 9) by averaging the information gains over the two possible outcomes ($S_{new} = \{0, 1\}$).

Alternative "no-update" model. We also test an alternative "no-update" model which is a lesioned version of the normative model. This lesioned variant skips the first step: it ignores the initial observation (β_{obs}, S_{obs}). Instead, it relies on its prior of competence $p(\theta)$ as its current belief when computing EIV. This prior is the same as the normative model. It still uses the same Ideal Observer likelihood for simulating updates from performance on a new question (β_{new}, S_{new}); it simply does not incorporate the initial observation before selecting a query. Comparing the normative model to this no-update model tests whether people adapt their search given the information presented.

Post-hoc heuristic model. In an exploratory analysis, we test a post-hoc model that was not pre-registered. This model predicts that, after observing a correct answer, participants will select harder queries proportionally to their difficulty. Conversely, when observing an incorrect answer, the model predicts the selection of easier questions. Note that this model does not implement a prior over competence, nor does it take into account the difficulty of the observed question. The predicted informational value is the difficulty of the query when observing a correct answer, and minus the difficulty when observing an incorrect answer.

4.2 Results

4.2.1 Confirmatory Analyses.

In this discrete choice task, participants had to choose a query among six possible options. Figure 6 plots the percentage of choice for each query (conditional on the observed question) compared to the normative model's predictions. Many queries within a given choice set have very similar EIV, so the model's predicted choice probabilities are only weakly differentiated. Under the normative model, EIV is highest for queries whose difficulty is near the estimated competence (i.e., where $p(S = 1) \approx 50\%$, as both a success or a failure would be informative).

To test H1, that participants are more likely to select more informative questions, a mixed multinomial logit model was estimated using the "mlogit" package (Croissant, 2020) with participants' choice of query as the dependent variable and the EIV for each query (computed with the normative model) as the independent variable. A participant-level random slope on EIV was added to allow the effect of EIV on choice to vary by individual. In accordance with our hypothesis, participants chose high-EIV queries at a rate significantly higher than chance ($\beta = 0.22 \pm 0.02$, $z = 9.42$, $p < .001$). In other words, one standard deviation increase in EIV increased the odds of selection by 24% on average. Participants were quite variable in how optimal they were in choosing the most diagnostic question, with an estimated standard deviation of the participant-specific EIV slopes of 0.32 ± 0.04 ($z = 8.59$, $p < .001$). As a descriptive complement to the mixed logit analysis, we computed the rank correlation between query EIV and choice percentage within each trial, then averaged across trials, obtaining a positive but relatively modest correlation (Spearman's $\rho = 0.34$).

As an illustration of these patterns, after a success on the easy item "What is the name of the first planet of the solar system?", the model determined that the most informative query should be of moderate difficulty because the observation provided little evidence about the individual's competence. The model, therefore, selected: "What are the only two planets of our solar system that do not possess any moons?" (EIV = 0.149). Some of the participants made the choice predicted by the model (21%), although the preferred choice was another question of similar difficulty: "What is the last planet of our solar system?" (EIV = 0.14; 33% of participants). After observing a failure on the same easy item, participants shifted to the easiest query to assess how incompetent the virtual individual might be: "What is the only liveable planet of our solar system?" (EIV = 0.082; 30% of participants). By contrast, the model favored the second easiest query as more informative

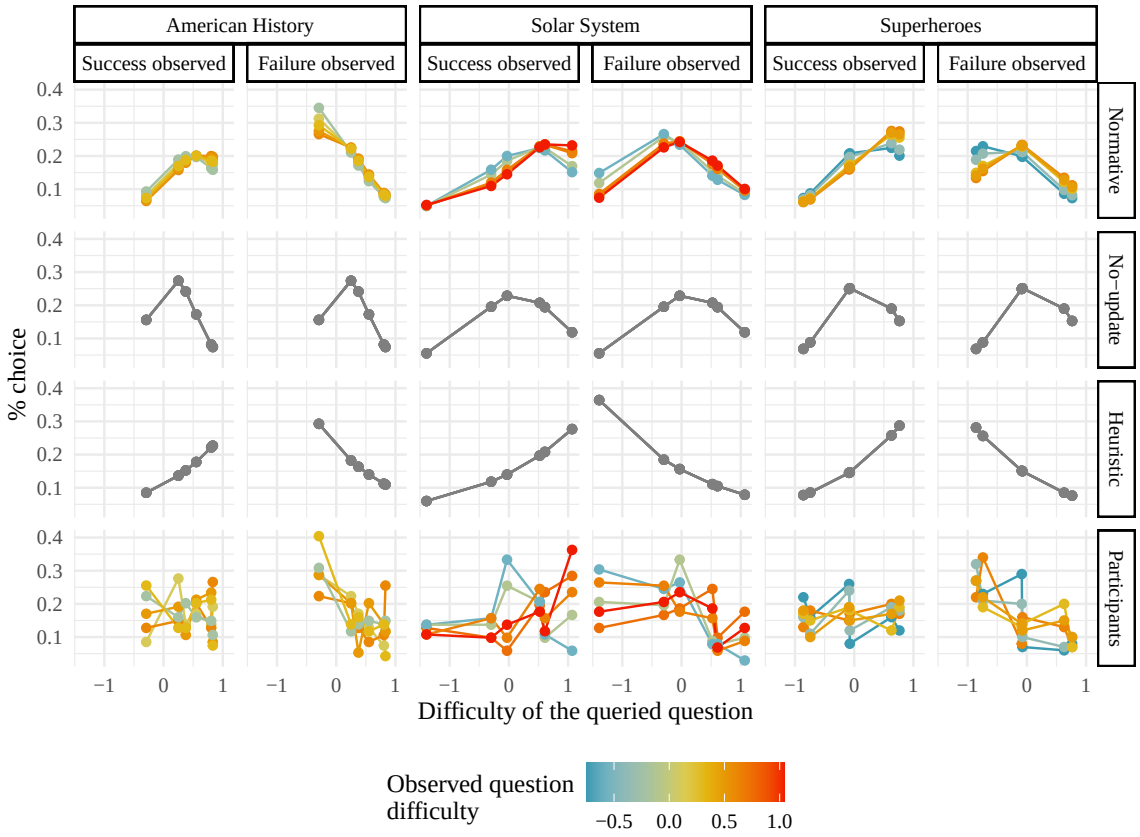

Figure 6. Percentage of choice predicted by the normative model, the no-update model, the heuristic model, and participants' average choice (bottom row), divided by theme and success. The query selections made after the same observation are grouped together by a line and colored in function of the difficulty of the observed question. Each data point corresponds to a pair of a query and an observed question, except for the no-update and heuristic model where the predictions are independent of the observed question. Transformation of the raw difficulty is described in section 2.1.4. To compare models' predictions with participants' choices, we transform EIV values to percentages of choice using a softmax transformation corresponding to a bounded-rational (noisy) policy ($\beta = 0.5$).

"What is Earth's only natural satellite?" (EIV = 0.127; 25% of participants). By contrast, after failure on a hard item ("Which planet has a hexagon shaped cloud formation on its north pole?"), participants and the model converged on a question of moderate difficulty "What is the last planet of our solar system?" (highest EIV = 0.154; 24% of participants).

To test that participants take into account the information previously presented about the individual they are evaluating (H2), we compared our normative model with an alternative "no-update" model which does not pay attention to the "observed question" (see Figure 6). We fitted a similar mixed multinomial logit model with the EIV predicted by the alternative "no-update" model. Participants' choice were not significantly correlated with this alternative estimate for EIV (β = -0.01 $\pm$ 0.02, z = -0.67, $p = 0.501$). Our normative model explains the data better ($BIC = 10501$) than the "no update" model ($BIC = 10586$).

4.2.2 Exploratory Analyses. We tested an alternative heuristic stipulating that when participants observe a correct answer they select harder questions and conversely, they select easier questions when observing an incorrect answer. We fitted another multinomial model with participants' choice as the dependent variable and the predictor was simply the difficulty of the query when observing a correct answer, and minus the difficulty when observing an incorrect answer. We found that participants' choices were significantly predicted by the heuristic ($\beta = 0.23 \pm 0.02$, $z = 10.62$, $p < .001$). As before, participants exhibited variation in how their choices followed the heuristic ($\beta = 0.20 \pm 0.04$, $z = 4.78$, $p < .001$). This heuristic model provided a marginally better account of the data than the normative model, as indicated by a lower BIC ($BIC = 10495$) and a slightly higher mean within-trial rank correlation (Spearman's $\rho = 0.37$).

In an exploratory analysis, we tested whether participants who observed a participant's answer to difficult questions subsequently picked more difficult questions to ask, a pattern predicted by the normative Bayesian model. A correct answer to a difficult question provides stronger evidence of high competence than a correct answer to an easy question, and failing a difficult question is less diagnostic of incompetence than failing an easy question. Therefore, in each case, the posterior over competence should be higher after observing a difficult compared to an easy question. When one's posterior estimate of competence is high, more difficult queries are more informative. We fitted a linear mixed effect model with participants as random intercepts and found that observing a difficult question was associated with selecting more difficult queries ($\beta = 0.13$, t(2,759.87) = 7.52, $p < .001$). This result highlights that, despite the heuristic model providing a similar fit to the data, there are aspects of the data that only the Bayesian model captures.

Study 3 shows that participants are able to flexibly query new information, in a way that can be predicted by an optimal search model based on information theory. Query selection increased with Expected Information Value, which provides convergent evidence that people use Bayesian inference both when estimating others' competence but also when deciding which new information to acquire.

5 General discussion and conclusion

The capacity to infer the competence and knowledge of others is a core component of social cognition. Studies 1 and 2 show that people can predict with a high degree of accuracy people's knowledgeability—more specifically, the conditional probability that a person knows the answer to a trivia question given that they have answered (or failed to answer) another trivia question. This confirms that even a single observation can drive major updates in competence assessment, if it is warranted by the diagnosticity of the observed performance (e.g. success on a very difficult item). These studies also provide evidence that inferences of competence are well approximated by a Bayesian model which optimally integrates novel information with prior expectations. By contrast, plausible heuristics, corresponding to lesioned versions of the Bayesian model, do not account as well for most participants' behavior. Ablation analyses show that assuming a non-uniform prior over competence, and stochasticity in the generative model ($\varepsilon > 0$), both contribute to the good fit of the Bayesian model. These two findings—accurate judgements, and reliance on normative inferences—are relatively independent, as applying non-normative heuristics sometimes results in accurate judgments, while normative inferences can produce biased answers when the generative model or the parameters are not well calibrated.

The same Bayesian framework can also explain how people seek information. Study 3 shows that participants are able to query new information in a way that can be predicted by an optimal search model based on information theory. The probability of selecting a query was influenced by the normative model Expected Information Value (a measure quantifying the amount of information gained for each candidate query). This result provides convergent evidence across distinct tasks that people's inferences of competence are consistent with a Bayesian model.

A large body of research has shown that humans systematically deviate from rational norms when reasoning under uncertainty (Kahneman, 2011; McDowell & Jacobs, 2017; Tversky & Kahneman, 1983). From this perspective, our results are surprising. Past research has identified several factors that improve statistical reasoning, like presenting information in a natural frequency format (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017). Most relevant to our findings, Krynski and Tenenbaum (2007) have shown that people are more likely to follow sound principles of statistical reasoning when they are given a clear causal model that explains the statistical relations they have to reason about (see also Tversky & Kahneman, 1980). In our experiments, we did not explicitly provide a causal model to participants, but their good performance provides some evidence that they may have spontaneously used a generative model of the relationship between competence and performance in order to draw inferences.

On the whole, performance was less in line with the Bayesian model in Study 3 (information search) than in Studies 1 and 2 (knowledge attribution). This difference might stem from the greater computational difficulty of Study 3, in which participants had to compare the potential outcomes of choosing each potential query weighted by the probability of a correct answer. Moreover, the Expected Information Values of each potential query were often similar. In such a situation, it is not surprising that a heuristic (asking difficult questions after observing a success, and easy questions after observing a failure) approximates

participants' behavior (on the rationality of using heuristics for computationally demanding tasks, see Bramley et al., 2015; Gigerenzer & Gaissmaier, 2011; Wu et al., 2017). Note that our focus was to characterize the cognitive mechanisms used during competence judgement and information-seeking at the computational level of analysis (Marr, 1982) and that we do not claim participants are necessarily implementing exact Bayesian inferences (for a discussion, see Griffiths et al., 2024).

The Bayesian model accurately captures inferences of knowledgeability, a specific facet of competence. We believe, however, that the model could be applied to other domains such as estimating technical skills (e.g. using the difficulty of some mental operations to predict numerical competence), or physical feats (e.g. using the effort required to lift an object of a given weight to estimate overall strength). As noted in the introduction, our computational framework is generalist as it does not make assumptions about the way difficulty is inferred. Estimating the difficulty of a trivia question probably recruits domain-specific mechanisms such as prior domain knowledge (Fastrich et al., 2018) or linguistic cues (e.g. length of the question, Boettcher, 2016), which are arguably distinct from inferences about other types of tasks (see Gweon et al., 2017). A possible extension of the model would be to look into situations of joint inference of competence and difficulty (Jansen et al., 2021; Leonard et al., 2019). While the current model assumes difficulty is known, a joint inference model would likely predict slower belief updating in ambiguous contexts, i.e., in which success or failure at the task is the only information about task difficulty. In such situations, failure could be attributed either to low competence or to high difficulty (Jones, 1989). The current models could also be extended to the estimation of competence or knowledge beyond the domain in which the performance has been observed—for instance, guessing someone's knowledge of astronomy, knowing they have correctly answered a question about superheroes.

The present studies have a number of limitations. As noted above, they only test participants' estimations of one form of competence (knowledgeability) and within specific domains. The information search task of Study 3 was arguably too difficult with the quality of many options being very hard to distinguish. Finally, we do not address the question of how participants are able to estimate so accurately the difficulty of trivia questions.

Our research opens new directions in understanding the fundamental mechanisms of social cognition. Building on prior literature establishing the importance of past accuracy and difficulty cues, the present research offers a systematic, computational account of how people utilize this information, tested again ecologically valid performance measure. We show that in a naturalistic paradigm, participants can use minimal—but reliable—cues to form accurate estimations of competence, in line with Bayesian principles. Future research could address the question of how participants weigh and integrate multiple cues of competence of varying diagnosticity.

References

Aboody, R., Davis, I., Dunham, Y., & Jara-Ettinger, J. (2025). People can infer the magnitude of other people's knowledge even when they cannot infer its contents. *Cognition*, *265*, 106236. https://doi.org/10.1016/j.cognition.2025.106236

Aboody, R., Lu, J., Denison, S., & Jara-Ettinger, J. (2025). Six-Year-Olds, but Not Younger Children, Consider the Probability of Being Right by Chance When Inferring Others' Knowledge. *Child Development*, *1*(12). https://doi.org/10.1111/cdev.14265

Altay, S., Majima, Y., & Mercier, H. (2020). It's my idea! Reputation management and idea appropriation. *Evolution and Human Behavior*, *41*(3), 235–243. https://doi.org/10.1016/j.evolhumbehav.2020.03.004

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018–1034. https://doi.org/10.1016/j.cognition.2007.12.008

Blasi, C. H., Bjorklund, D. F., & Soler, M. R. (2015). Cognitive Cues are More Compelling than Facial Cues in Determining Adults' Reactions towards Young Children. *Evolutionary Psychology*, *13*(2), 511–530. https://doi.org/10.1177/147470491501300212

Boettcher, E. (2016). *Predicting the difficulty of trivia questions using text features* [Master's thesis]. University of North Carolina at Chapel Hill.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731. https://doi.org/10.1037/xlm0000061

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2020). Contributions of Nonverbal Cues to the Accurate Judgment of Personality Traits. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment* (1st ed., pp. 194–218). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.13

Breit, M., Scherrer, V., Tucker-Drob, E. M., & Preckel, F. (2024). The Stability of Cognitive Abilities: A Meta-Analytic Review of Longitudinal Studies. *Psychological Bulletin*, *150*(4), 399–439. https://doi.org/10.1037/bul0000425

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Castelain, T., Bernard, S., & Mercier, H. (2018). Evidence that Two-Year-Old Children are Sensitive to Information Presented in Arguments. *Infancy*, *23*(1), 124–135. https://doi.org/10.1111/infa.12202

Castelli, L., Carraro, L., Ghitti, C., & Pastore, M. (2009). The effects of perceived competence and sociability on electoral outcomes. *Journal of Experimental Social Psychology*, *45*(5), 1152–1155. https://doi.org/10.1016/j.jesp.2009.06.018

Clegg, J. M., Kurkul, K. E., & Corriveau, K. H. (2019). Trust me, I'm a competent expert: Developmental differences in children's use of an expert's explanation quality

to infer trustworthiness. *Journal of Experimental Child Psychology*, *188*, 104670. https://doi.org/10.1016/j.jecp.2019.104670

Corriveau, K., & Harris, P. L. (2009a). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426–437. https://doi.org/10.1111/j.1467-7687.2008.00792.x

Corriveau, K., & Harris, P. L. (2009b). Preschoolers continue to trust a more accurate informant 1 week after exposure to accuracy information. *Developmental Science*, *12*(1), 188–193. https://doi.org/10.1111/j.1467-7687.2008.00763.x

Corriveau, K., Kinzler, K. D., & Harris, P. L. (2013). Accuracy trumps accent in children's endorsement of object labels. *Developmental Psychology*, *49*(3), 470–479. https://doi.org/10.1037/a0030604

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73. https://doi.org/10.1016/0010-0277(95)00664-8

Croissant, Y. (2020). Estimation of Random Utility Models in R: The mlogit Package. *Journal of Statistical Software*, *95*, 1–41. https://doi.org/10.18637/jss.v095.i11

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–648. https://doi.org/10.1037/0022-3514.92.4.631

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98. https://doi.org/10.1016/j.riob.2011.10.004

Davis, Z. J., Allen, K. R., Kleiman-Weiner, M., Jara-Ettinger, J., & Gerstenberg, T. (2025). Inference from social evaluation. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000445

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455–476. https://doi.org/10.1037/rev0000175

Dubourg, E., Morin, O., & Mercier, H. (2025). Using the Nested Structure of Knowledge to Infer What Others Know. *Psychological Science*, *36*(6), 443–450. https://doi.org/10.1177/09567976251339633

Eisenbruch, A. B., Smith, K. M., Workman, C. I., von Rueden, C., & Apicella, C. L. (2024). US adults accurately assess Hadza and Tsimane men's hunting ability from a single face photograph. *Evolution and Human Behavior*, *45*(4), 106598. https://doi.org/10.1016/j.evolhumbehav.2024.106598

Elashi, F. B., & Mills, C. M. (2014). Do children trust based on group membership or prior accuracy? The role of novel group membership in children's trust decisions. *Journal of Experimental Child Psychology*, *128*, 88–104. https://doi.org/10.1016/j.jecp.2014.07.003

Fastrich, G. M., Kerr, T., Castel, A. D., & Murayama, K. (2018). The role of interest in memory for trivia questions: An investigation with a large-scale database. *Motivation Science*, *4*(3), 227–250. https://doi.org/10.1037/mot0000087

Fousiani, K., Sypes, C., & Armenta, B. M. (2023). Applying for remote jobs? You'd better be competent! Teleworking turns recruiters attention to candidate competence over warmth-related skills. *Personality and Individual Differences*, *204*, 112063. https://doi.org/10.1016/j.paid.2022.112063

Garfield, Z. H., von Rueden, C. R., & Hagen, E. H. (2025). The Multi-Capital Leadership Theory. *Human Nature*. https://doi.org/10.1007/s12110-025-09503-y

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781. https://doi.org/10.1016/j.visres.2010.09.027

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(Volume 62, 2011), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (Eds.). (2024). *Bayesian Models of Cognition: Reverse Engineering the Mind*. The MIT Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *39*.

Gweon, H., Shafto, P., & Schulz, L. (2014). Children consider prior knowledge and the cost of information both in learning from and teaching others. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive Foundations of Learning from Testimony. *Annual Review of Psychology*, *69*(1), 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Himmelstein, M., Atanasov, P., & Budescu, D. V. (2021). Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment and Decision Making*, *16*(2), 323–362. https://doi.org/10.1017/S1930297500008597

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26.

Jaeger, B., Evans, A. M., Stel, M., & Van Beest, I. (2022). Understanding the role of faces in person perception: Increased reliance on facial appearance when judging sociability. *Journal of Experimental Social Psychology*, *100*, 104288. https://doi.org/10.1016/j.jesp.2022.104288

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763. https://doi.org/10.1038/s41562-021-01057-0

Jara-Ettinger, J., & Gweon, H. (2017). Minimal covariation data support future one-shot inferences about unobservable properties of novel agents. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *39*(0).

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as

a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334. https://doi.org/10.1016/j.cogpsych.2020.101334

Jones, E. E. (1989). The Framing of Competence. *Personality and Social Psychology Bulletin*, *15*(4), 477–492. https://doi.org/10.1177/0146167289154001

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi.org/10.1037/0033-295X.94.2.211

Klopfenstein, A., & Mercier, H. (2025). Explaining is not enough: Appealing explanations should also be surprising. *Psychonomic Bulletin & Review*.

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in Testimony: Children's Use of True and False Statements. *Psychological Science*, *15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430–450. https://doi.org/10.1037/0096-3445.136.3.430

Kryven, M., Ullman, T. D., Cowan, W., & Tenenbaum, J. B. (2021). Plans or Outcomes: How Do We Attribute Intelligence to Others? *Cognitive Science*, *45*(9), e13041. https://doi.org/10.1111/cogs.13041

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior*, *32*(1), 4–14. https://doi.org/10.3758/BF03196002

Lane, J. D., Wellman, H. M., & Gelman, S. A. (2013). Informants' Traits Weigh Heavily in Young Children's Trust in Testimony and in Their Epistemic Inferences. *Child Development*, *84*(4), 1253–1268. https://doi.org/10.1111/cdev.12029

Leonard, J. A., Bennet-Pierre, G., & Gweon, H. (2019). Who is better? Preschoolers infer relative competence based on efficiency of process and quality of outcome. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Liefgreen, A., Pilditch, T., & Lagnado, D. (2020). Strategies for selecting and evaluating information. *Cognitive Psychology*, *123*. https://doi.org/10.1016/j.cogpsych.2020.101332

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The Child as Econometrician: A Rational Model of Preference Understanding in Children. *PLOS ONE*, *9*(3), e92160. https://doi.org/10.1371/journal.pone.0092160

Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four- and 5-Year-Olds Infer Differences in Relative Ability and Appropriately Allocate Roles to Achieve Cooperative, Competitive, and Prosocial Goals. *Open Mind*, *2*(2), 72–85. https://doi.org/10.1162/opmi_a_00019

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380. https://doi.org/10.1016/j.cognition.2009.05.012

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on

Bayesian reasoning. *Psychological Bulletin*, *143*(12), 1273–1312. https://doi.org/10.1037/bul0000126

Menegatti, M., Pireddu, S., Crocetti, E., Moscatelli, S., & Rubini, M. (2021). The Ginevra de' Benci Effect: Competence, Morality, and Attractiveness Inferred From Faces Predict Hiring Decisions for Women. *Frontiers in Psychology*, *12*, 658424. https://doi.org/10.3389/fpsyg.2021.658424

Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton University Press.

Najar, A., Bonnet, E., Bahrami, B., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLOS Biology*, *18*(12), e3001028. https://doi.org/10.1371/journal.pbio.3001028

Nelson, J. D. (2005). Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain. *Psychological Review*, *112*(4), 979–999. https://doi.org/10.1037/0033-295X.112.4.979

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216–1226. https://doi.org/10.1037/0012-1649.43.5.1216

Quillien, T. (2023). Rational information search in welfare-tradeoff cognition. *Cognition*, *231*, 105317. https://doi.org/10.1016/j.cognition.2022.105317

Quillien, T., & Taylor-Davies, M. (2025). *Factive mindreading reflects the optimal use of limited cognitive resources*. PsyArXiv. https://doi.org/10.31234/osf.io/wzbcs_v1

Quillien, T., Tooby, J., & Cosmides, L. (2023). Rational inferences about social valuation. *Cognition*, *239*, 105566. https://doi.org/10.1016/j.cognition.2023.105566

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests* (pp. xiii, 184). Nielsen & Lydiche.

Reyes-Jaquez, B., & Echols, C. H. (2013). Developmental differences in the relative weighing of informants' social attributes. *Developmental Psychology*, *49*(3), 602–613. https://doi.org/10.1037/a0031674

Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: The hidden costs to women of a kinder, gentler image of middle managers. *Journal of Personality and Social Psychology*, *77*(5), 1004–1010. https://doi.org/10.1037/0022-3514.77.5.1004

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*(3), 436–447. https://doi.org/10.1111/j.1467-7687.2012.01135.x

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Taylor, M. G. (2013). Gender influences on children's selective trust of adult testimony. *Journal of Experimental Child Psychology*, *115*(4), 672–690. https://doi.org/10.1016/j.

jecp.2013.04.003

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration Trumps Confidence as a Basis for Witness Credibility. *Psychological Science*, *18*(1), 46–50. https://doi.org/10.1111/j.1467-9280.2007.01847.x

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of Competence from Faces Predict Election Outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. In *Advances in Experimental Social Psychology* (Vol. 63, pp. 189–245). Elsevier. https://doi.org/10.1016/bs.aesp.2020.11.004

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*(Volume 66, 2015), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Török, G., Swaboda, N., & Ruggeri, A. (2023). Smart or Just Lucky? Inferring Question-Asking Competence From Strategies' Efficiency Versus Effectiveness. *Developmental Psychology*, *59*(6), 1136–1152. https://doi.org/10.1037/dev0001516

Tsividis, P., Gershman, S., Tenenbaum, J., & Schulz, L. (2014). Information selection in noisy environments with large action spaces. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*.

Turpin, M. H., Kara-Yakoubian, M., Walker, A. C., Walker, H. E. K., Fugelsang, J. A., & Stolz, J. A. (2021). Bullshit Ability as an Honest Signal of Intelligence. *Evolutionary Psychology*, *19*(2), 14747049211000317. https://doi.org/10.1177/14747049211000317

Tversky, A., & Kahneman, D. (1980). Causal Schemas in Judgments Under Uncertainty. In M. Fishbein (Ed.), *Progress in Social Psychology* (1st Edition, pp. 49–72). Psychology Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293. https://doi.org/10.1037/0033-295X.90.4.293

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2017). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior*, *38*(1), 9–17. https://doi.org/10.1016/j.evolhumbehav.2016.06.002

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1274–1297. https://doi.org/10.1037/xlm0000374

Xiang, Y., Gershman, S. J., & Gerstenberg, T. (2026). A signaling theory of self-handicapping. *Cognition*, *266*, 106288. https://doi.org/10.1016/j.cognition.2025.106288

Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.