Inferring Arithmetic Skill from Speed and Accuracy

Marius Mercier[1,†,*], **Ruby de Lanerolle**[2,†], **Olivier Morin**[1], **Tadeg Quillien**[3] **& Hugo Mercier**[1]
[1]Institut Jean Nicod, Département d'études cognitives,
École normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France
[2]Department of Experimental Psychology, UCL, United Kingdom
[3]Department of Psychology, University of Edinburgh, United Kingdom
[†]These authors contributed equally. [*]Corresponding author: mariusmercier1@gmail.com

Abstract

People routinely infer others' competence under uncertainty, often relying on cues such as task difficulty and past accuracy. An emerging body of research suggests that people approximate Bayesian inference when doing so. We extend these results by testing whether people can infer others' numerical ability in a way that is consistent with a rational Bayesian model. In Study 1, we find that participants accurately predict the arithmetic performance of another individual from information about their past performance. Computational modeling shows that participants' inferences are better described by Bayesian processes than by plausible heuristics. Study 2 introduces a modified paradigm, in which participants are told about both past performance and time taken to solve problems. We find that, although participants are quite accurate in their predictions, they do not seem to take into account information about speed.

Keywords: Computational Modeling; Social Cognition; Competence; Numerical Cognition

Introduction

Humans routinely make decisions that depend on judgments of others' competence: whom to ask for advice, whom to hire, or whom to cooperate with (Cuddy et al., 2011; Harris et al., 2018; Xiang et al., 2023). Yet competence is not directly observable. Observers must infer it from sparse and noisy cues, and even apparently informative outcomes can be ambiguous (Jones, 1989; Rasch, 1960). Consider choosing between two teachers after observing a single performance from each of them: one solves an easy problem, the other fails a hard one. Despite the contrast between success and failure, neither observation is very diagnostic, because easy successes and predictable failures are both common.

A growing body of work suggests that people nonetheless make competence inferences that approximate rational probabilistic reasoning (Aboody et al., 2025; Baker et al., 2017; Davis et al., 2025; Jansen et al., 2021; Jara-Ettinger & Gweon, 2017; Xiang et al., 2026). In particular, Mercier et al. (2026) used a naturalistic trivia paradigm to show that participants infer what others know from minimal evidence (a single success or failure), by combining the observed outcome with prior expectations about item difficulty and the average competence of the population in a domain. In this framework, competence judgments integrate (i) the diagnosticity of the observed performance—success on a hard item is more informative than success on an easy one—and (ii) prior expectations about the individual's competence.

However, previous research primarily focuses on knowledgeability, a specific facet of competence (Aboody et al., 2025; Dubourg et al., 2025; Mercier et al., 2026). This leaves open a key question about generality: do people apply similar inference principles to skill-based competence? Knowledge-based and skill-based competence rely on distinct generative processes. Inferring what someone knows requires representing their cumulative exposure to information, which can be seen as a gradient of interest or exposure. Inferring what someone can do, by contrast, requires representing a gradient of ability, typically attributed to some combination of aptitude and accumulated practice, that may or may not be sufficient to handle the problem at hand. This composite structure may complicate skill inference relative to knowledge inference, since the same observed performance can reflect different mixtures of aptitude and practice. Whether the same Bayesian inferential logic extends across these two generative processes is an open question. We test this generality using mental arithmetic, a domain with a well-validated difficulty gradient (Dehaene, 2011) that lends itself to time-based manipulations. This lets us test whether observers integrate speed alongside accuracy.

Our contribution is twofold. First, in Study 1 we adapt the trivia paradigm of Dubourg et al. (2025) to arithmetic and test whether participants' judgments about another person's numerical ability are consistent with a Bayesian model that links latent competence to success probabilities across item difficulties. We also compare this normative account to heuristic baselines previously considered in the trivia domain (Mercier et al., 2026).

Second, in Study 2 we test whether observers additionally use speed as a cue to competence when accuracy information is held constant. While speed is often cited as a marker of arithmetic fluency (Cheng et al., 2021; Roy et al., 2025), it is theoretically ambiguous: faster responses may signal efficiency, but they can also reflect lower response caution; for complex reasoning tasks, competent individuals often take longer to verify their answers (see Berke et al., 2023; Goldhammer & Klein Entink, 2011; Hedge et al., 2018; Kruger et al., 2004; Richardson & Keil, 2022). We therefore ask whether people's competence judgments track the true diagnostic value of response time in our task.

Computational Model

Inferring competence from performance requires accounting for uncertainty: people can guess, slip, or make occasional errors even when skilled (Jones, 1989). We formalize competence inference using the Bayesian framework developed in Mercier et al. (2026) and closely related to item-level models of ability (Jansen et al., 2021). Observers are assumed to represent a latent competence parameter θ for an individual and a known item difficulty parameter β for each task. After observing whether an individual succeeds ($S = 1$) or fails ($S = 0$) on an item of difficulty β, observers update beliefs about θ via Bayes' rule:

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

The likelihood $p(S \mid \beta, \theta)$ links competence and difficulty through a psychometric function (with an Item Response Theory mapping, Rasch, 1960), allowing outcomes to be stochastic rather than deterministic. Given the posterior over θ, the observer can then predict performance on new items by marginalizing over uncertainty in competence. The model architecture (Bayesian updating with a Rasch-style likelihood, together with the heuristic baselines) follows Mercier et al. (2026); our contribution is to test this framework in a skill-based domain. The full specification of the Bayesian model and heuristic models is provided in the Methods for Study 1.

Study 1

In this study, we use a paradigm similar to Mercier et al. (2026) to investigate the inferences people make about individuals who provide either a correct or incorrect answer to a math question. We predicted that participants' behavior would be best captured by the optimal Bayesian Model (henceforth "the Main Model"). More precisely, we put forward the following hypotheses[1]:

H1. Difficulty judgments. For different questions, participants can accurately assess the probability of reaching a correct answer within the time limit.

H2. Competence inferences from a single outcome. On average, participants' predictions of the probabilities of accurate answers on different questions, given the observation of a failure/success on a particular question, will be significantly correlated with true conditional probabilities.

H3. Model accuracy. The predictions of the Main Model are positively correlated with participants' average predictions.

H4. Model comparison. When fitted to the aggregated dataset, the Main Model better captures participants' behavior than heuristic models.

[1]The procedures, data collection, analysis plan and the models were pre-registered for Study 1 and 2: https://osf.io/n9erf/files/osfstorage?view_only=12991dcd2fba4fe2b9a1a77339d6cb2f. Data and code are openly available: https://github.com/mariusmercier/paper_numeric_competence/tree/cogsci-2026

Methods

Participants 300 U.S. participants were recruited via Prolific. We applied pre-registered exclusions for inattentiveness/LLM use (N = 2), suspected use of external calculation aids (N = 35), and manipulation noncompliance (changing the pre-ticked success/failure response; N = 39), leaving a sample of 224 (120 women, 104 men, M_{Age} = 43.4, SD_{Age} = 14.8)

Procedure This experiment replicates the paradigm of Mercier et al. (2026), but with different stimuli and a time constraint. Participants were asked to complete a task in two phases. In the judgment phase, they were asked to evaluate five virtual agents on a set of fifteen mathematical questions. Participants were told that in order to help them, they would have access to the performance of the virtual agent on one of the questions. For each trial, participants were assigned to a Success/Failure condition: each virtual agent correctly ("Success" condition) or incorrectly ("Failure" condition) answered the question under 30 seconds. The specific question shown was randomly drawn from a stock of 15 questions. Only the question, and whether the virtual agent got it right or not, was provided to the participants, but not the specific answer. We call this item the "observed question."

Showing only the outcome (rather than the agent's specific answer) serves two purposes. First, it isolates competence inference from the participant's own ability to evaluate that answer, which would otherwise confound the inferred competence of the agent with the competence of the evaluator. Second, it keeps the inference computationally tractable: a binary success/failure outcome maps cleanly onto the likelihood term in a Bayesian model, whereas reasoning over the space of possible wrong answers would require a richer generative model of arithmetic errors.

Participants were then asked to evaluate whether the virtual agent succeeded or failed to answer the other 14 questions from the quiz within the time limit. We call each of these items a "new question." Observed-new question pairs are thus, for each observed question, the 14 predictions on new questions. Participants also had to estimate how many people, out of 100, correctly answered the observed question within 30 seconds.

In the second phase, participants had to complete the 15 maths questions, and one additional question designed to detect participants using external tools. The order was randomly determined. They were instructed that they had to answer each question within 30 seconds. At any time, once they had entered an answer, they could click Submit to proceed. At 15 seconds and 25 seconds, time warnings appeared. If after 30 seconds a participant had not submitted an answer, the survey automatically proceeded to the next question and the timer restarted.

Materials The arithmetic stimuli were designed to establish a difficulty gradient following principles specified in Dehaene (2011). We manipulated five parameters to vary complexity: (1) answer magnitude (number of digits); (2) semantic roundness (operands ending in zeros vs. "sharp" non-zero

integers); (3) operand count; (4) carry-over frequency (aggregate sums >10); and (5) placeholder tracking (alignment required for variable-length operands). For example, the problem 100,034,090 + 1,023 represents a high-difficulty item as it requires placeholder tracking, mixed sharp/round operands, and a carry-over operation.

Computational Models We used the Bayesian/IRT framework from Mercier et al. (2026). Each question is assigned a difficulty score β by averaging judged difficulty ratings (divided by 100) and applying a logit transform to map scores from (0,1) to an unbounded scale.[2] Competence θ has a normal prior $\theta \sim \mathcal{N}(\mu, \sigma^2)$, and success is stochastic under a normal-ogive Rasch likelihood:

$$p(S = 1 \mid \beta, \theta) = \Phi\left(\frac{\theta - \beta}{\varepsilon}\right) \tag{2}$$

After observing $(\beta_{\text{obs}}, S_{\text{obs}})$, the posterior is proportional to $p(S_{\text{obs}} \mid \beta_{\text{obs}}, \theta)\, p(\theta)$ (Equation 1). The predicted success on a new item is

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

The Bayesian model has four free parameters: $\mu, \sigma^2, \varepsilon$, and a softmax temperature τ that maps probability estimates to binary responses. We compare this model to the same two heuristic baselines (Threshold, Anchor) described in Mercier et al. (2026).

The Threshold heuristic assumes a deterministic rule: after a success, competence is at least the observed difficulty; after a failure, competence is at most that difficulty. It treats all values as equally likely, then predicts success on a new item by the posterior mass above its difficulty (with τ governing response noise). The Anchor heuristic collapses inference to a single point estimate: after success, competence is set a fixed distance Δ above the observed difficulty; after failure, Δ below. It then predicts success deterministically for items easier than that anchor (again with τ for response noise).

Results

Difficulty judgments (H1). Participants' perceived difficulty for each item closely tracked objective difficulty (i.e. the percentage of participants not being able to solve an item within the time limit in the questionnaire phase). A linear mixed-effects model predicting perceived difficulty from objective difficulty at the trial-level yielded a strong positive association ($\beta = 0.45$, $t(911.69) = 22.18$, $p < .001$; see Figure 1A).

[2]This transformation is inspired by the way difficulty is operationalized in frameworks like Item Response Theory (Rasch, 1960). The qlogis transform is given by: $qlogis(p) = log(p/(1-p))$.

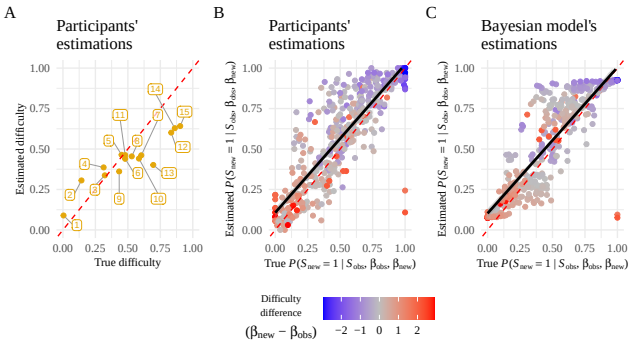

Figure 1: (**A**) Each question's estimated difficulty as a function of its true difficulty. Numbers refer to the question ID (see Materials). The red dashed line represents a perfectly accurate estimation. (**B**) Participants' estimated probability of answering a new question correctly, conditioned on the performance on the observed question, as a function of true conditional probabilities. (**C**) The Bayesian model's estimated conditional probabilities as a function of true conditional probabilities. For (**B**) and (**C**), each data point corresponds to the average prediction for an observed-new question pair. The red line represents a perfectly accurate prediction, while the dark line represents the fit of the linear model.

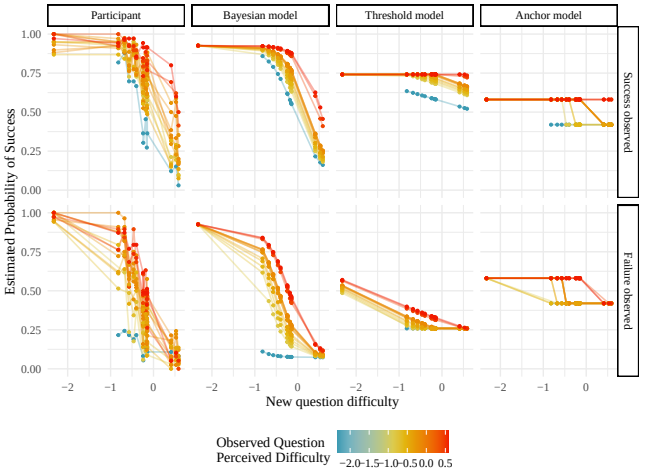

Figure 2: Model comparison for Study 1. Each data point corresponds to the average prediction for an observed-new question pair. The left column shows participants' average predictions, while the remaining columns show predictions from the Bayesian model, Threshold model, and Anchor model respectively. The top row shows trials where success was observed, and the bottom row shows trials where failure was observed. Lines connect predictions for the same observed question across different evaluated questions, colored by the observed question's perceived difficulty.

Competence inferences from a single outcome (H2). Participants used the observed success/failure to make predictions about performance on other items. We compare these predictions to the true conditional probabilities from the participants' Quiz performance: for each observed-new question pair, the proportion of participants who solved the new question given a success/failure on the observed question (matching the structure of the prediction task). Averaged predictions for each observed-new question pair were strongly correlated with the true conditional probabilities ($\beta = 0.86$, $t(418) = 34.72$, $p < .001$, $R^2_{\text{adj}} = 0.74$; Figure 1B).

Model accuracy and comparison (H3–H4). The Bayesian model reproduced participants' average conditional predictions across item pairs (Main Model: $\beta = 0.94$, $t(418) = 58.29$, $p < .001$, $R^2_{\text{adj}} = 0.89$), outperforming both heuristic models (Threshold: $R^2_{\text{adj}} = 0.39$; Anchor: $R^2_{\text{adj}} = 0.59$; see Figure 2). Model selection using BIC favored the Bayesian model over Threshold and Anchor ($BIC_{\text{Bayes}} = 16063.72$, $BIC_{\text{Threshold}} = 19411.92$, $BIC_{\text{Anchor}} = 18259.96$), providing strong evidence for the Bayesian account.

Robustness analyses. To test whether accurate competence inference requires high arithmetic ability, we replicated H1–H2 in the bottom 30% of performers on the arithmetic test. Even in this subsample, perceived difficulty tracked objective difficulty ($\beta = 0.93$, $t(303) = 44.51$, $p < .001$), and average conditional predictions remained strongly correlated with true conditional probabilities ($\beta = 0.85$, $t(448) = 34.03$, $p < .001$, $R^2_{\text{adj}} = 0.72$).

Study 2

Study 2 tests whether observers incorporate information about an answer's speed, beyond its accuracy, as a cue to numerical competence. Theoretically, response time is an ambiguous signal. Fast correct answers can reflect fluency but also shallow processing or guessing; long response times can reflect struggle and low automaticity but also deliberation, caution, or metacognitive monitoring (Berke et al., 2023; Goldhammer & Klein Entink, 2011; Hedge et al., 2018; Kruger et al., 2004; Richardson & Keil, 2022).

To isolate the diagnostic value of speed, we introduced an exogenous time constraint. Unlike spontaneous response times, which are conflated with an agent's internal speed-accuracy trade-offs, a strict time limit places a hard bound on performance capability. We manipulated this constraint by informing participants that the target agent had to answer within either a Fast (20 s) or Slow (40 s) time-limit.

This manipulation generates distinct predictions. A success under the Fast limit is stronger evidence of high competence than a success under the Slow limit, as it rules out the possibility that the agent needed the extra time to solve the problem. Conversely, a failure under the Fast limit is weaker evidence of low competence than a failure under the Slow limit, as the former might merely reflect insufficient time rather than a lack of skill. Therefore, if observers are rational updaters, observing performance in the Fast condition should, all else being equal, lead to higher competence estimates than observing performance in the Slow condition.

More precisely, we pre-registered the following hypotheses:

H1. Difficulty judgments. Participants accurately assess the difficulty of correctly answering the observed questions.

H2a. Fast accurate answers are competence cues. For a given pair of Quiz 1 and Quiz 2 questions, among participants who correctly answered the Quiz 2 question, those who answered it under a 20-second time limit have a higher chance of answering the Quiz 1 question than those who answered under a 40-second time limit.

H2b. Fast inaccurate answers are competence cues. For a given pair of Quiz 1 and Quiz 2 questions, among participants who failed the Quiz 2 question, those who failed under a 20-second time limit have a higher chance of answering the Quiz 1 question than those who failed under a 40-second time limit.

H3. Competence inferences from a single outcome. On average, participants' predictions of the probabilities of success on different evaluated questions, given success or failure within the 20-second or 40-second time limit on an observed question, are significantly correlated with true conditional probabilities.

H4a. Participants treat faster accurate answers as competence cues. For a given pair of Quiz 1 and Quiz 2 questions, participants' predictions of the probabilities of success on different evaluated questions will be higher for correctly answered observed questions under 20 seconds rather than under 40 seconds.

H4b. Participants treat faster inaccurate answers as competence cues. For a given pair of Quiz 1 and Quiz 2 questions, participants' predictions of the probabilities of success on different evaluated questions will be higher for incorrectly answered observed questions under 20 seconds rather than under 40 seconds.

Methods

Participants We recruited 400 U.S. participants via Prolific Academic. As in Study 1, we excluded participants who passed the cheat-test item (N = 25) or reported using calculation aids (N = 18); no one failed the attention check (N = 0). We additionally implemented bot-detection measures (N = 7). The final sample comprised N = 354 (224 women; $M_{Age} = 41.7$, $SD_{Age} = 12.7$).

Behavioral task As in Study 1, participants completed a Quiz Phase (to measure performance and calibrate difficulty) and a Judgment Phase (to predict a target's performance on other items given one observed item). Study 2 differed in two ways: (i) the Quiz Phase was completed first, and (ii) the observed performance in the Judgment Phase included a time-limit cue (Fast vs. Slow). The Quiz-first ordering was chosen so that participants would have direct experience of the time limit before being asked to reason about it.

Participants completed three arithmetic quizzes of 5 items each. Quiz 1 was always Slow (40 seconds per item). Quiz 2 was randomly assigned across participants to be Fast (20 seconds) or Slow (40 seconds), and Quiz 3 was counterbalanced so that all participants experienced both constraints. Each quiz used a distinct item set (constant across participants), with items randomized within quiz; correct answers were never shown. As in Study 1, participants were instructed to use mental calculation only, and a cheat-test item (used for exclusions) appeared at the end of Quiz 1.

The Judgment Phase followed Study 1, except that the observed outcome was paired with a time limit (e.g. "the target answered correctly within 20 s" vs. "within 40 s"). Participants evaluated 5 virtual agents. For each agent, they observed one Quiz 2 item and outcome (success/failure) with its time limit, then predicted whether the agent would succeed on each of the 5 Quiz 1 items under that same time limit. Participants also provided a perceived-difficulty judgment for the observed item that incorporated the time limit (e.g. "Out of 100 participants, how many would solve this item within 20 s?"). Observed items, outcomes, and time limits were randomly assigned and balanced across participants.

Materials Materials consisted of three sets of five arithmetic items (Quizzes 1–3). In the Judgment Phase, Quiz 1 items served as evaluated questions and Quiz 2 items as observed questions.

Each quiz was constructed to span the same difficulty gradient. Quiz 1 comprised five items selected from Study 1 based on empirical accuracy rates, yielding roughly evenly spaced difficulty levels. Quiz 2 and Quiz 3 were constructed to match these levels by aligning core determinants of arithmetic difficulty (Dehaene, 2011) (e.g. number of digits, operand "roundness," number of carries, number of operands, and placeholder tracking).

Results

Difficulty judgments (H1). We first tested whether participants could assess the difficulty of correctly answering the observed questions. Difficulty judgments were collected for Quiz 2 items only (judgments on Quiz 1 items were already collected in Study 1). Because participants encountered both time limits equally often across the Quiz and Judgment phases, we did not subdivide this analysis by time condition. A linear mixed-effects model with random intercepts and slopes for participants showed that perceived difficulty tracked objective difficulty ($\beta = 0.39$, $t(313.06) = 16.85$, $p < .001$). Note that difficulty judgments were collected after the Quiz Phase in Study 2 but before it in Study 1. The replication of H1 across studies therefore argues against the possibility that prior exposure to the items drives difficulty calibration.

Is response time objectively diagnostic? (H2a–H2b). Next, we tested whether the time-limit condition attached to the observed Quiz 2 performance provided information about competence beyond accuracy. We constructed all pairs of Quiz 1 (evaluated) and Quiz 2 (observed) items from participants' quiz performance (e.g. Q1.1–Q2.1, Q1.1–Q2.2, etc.), mirroring the structure of the Judgment Phase. We estimated two mixed-effects logistic regressions: one restricted to trials where the Quiz 2 item was solved (testing H2a), and one restricted to trials where the Quiz 2 item was failed (testing H2b). In both models, the outcome was whether the Quiz 1 item was answered correctly, and the predictor was the observed time limit in Quiz 2 (Fast: 20 s vs. Slow: 40 s), with random intercepts for participants and for the item-pair (thus comparing Fast vs. Slow within the same Q1–Q2 pair, controlling for accuracy).

Consistent with H2a, among participants who solved the Quiz 2 item, those who solved it under the Fast limit were more likely to also solve the Quiz 1 item than those who solved it under the Slow limit (H2a: $b = 0.51$, $z = 2.15$, $p = 0.031$). Consistent with H2b, among participants who failed the Quiz 2 item, those who failed under the Fast limit were likewise more likely to solve the Quiz 1 item than those who failed under the Slow limit (H2b: $b = 0.71$, $z = 2.63$, $p = 0.009$). Thus, holding accuracy constant, the Fast condition was (weakly) informative of higher competence in the ground truth.

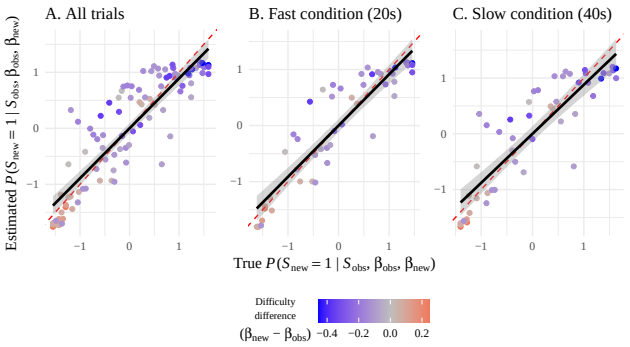

Figure 3: Estimated Conditional Probability versus True Conditional Probability in Study 2. Each data point is a pair of two questions. The red line represents the optimal prediction while the dark line represents the fit of the linear models. (**A**) shows all trials combined, (**B**) shows trials in the Fast condition (20-second time limit), and (**C**) shows trials in the Slow condition (40-second time limit). Points are colored by the difficulty difference between the evaluated and observed questions.

Competence inferences from a single outcome (H3). We then tested whether participants' judgments about the target's performance on Quiz 1 items tracked true conditional probabilities when conditioning on both the observed outcome (success/failure) and the time limit. Replicating Study 1's analysis, averaged predictions were strongly correlated with true conditional probabilities computed from quiz performance (H3: $\beta = 0.89$, $t(98) = 19.47$, $p < .001$, $R^2_{adj} = 0.79$; Figure 3).

This relationship held across both time-limit conditions (Figure 3B–C).

Do participants use the time-limit cue? (H4a–H4b). Finally, we tested whether participants incorporated the time-limit information when making predictions. We estimated models parallel to those used for H2a–H2b, but with participants' binary predictions (rather than true Quiz 1 performance) as the dependent variable, again controlling for item-pair and restricting analyses to observed success (H4a) or observed failure (H4b). Participants' predictions did not reliably differ between the Fast and Slow conditions, either after observing a correct answer (H4a: $b = 0.17$, $z = 1.53$, $p = 0.125$) or after observing an incorrect answer (H4b: $b = 0.15$, $z = 1.62$, $p = 0.104$). Thus, although the time-limit manipulation carried some diagnostic information in the performance data (H2a–H2b), participants did not appear to use it in their judgments.

Robustness analyses. We further tested whether H1 and H3 held for participants with minimal arithmetic ability (bottom 30% of scores in the Quiz Phase). These participants still accurately tracked objective difficulty ($\beta = 0.32$, $t(88.04) = 8.16$, $p < .001$) and their average predictions remained strongly correlated with true conditional probabilities ($\beta = 0.88$, $t(98) = 18.73$, $p < .001$, $R^2_{\text{adj}} = 0.78$). We further replicated H3 with a trial level analysis and found that participants' predictions also covaried with true conditional probabilities ($b = 2.32$, $z = 31.58$, $p < .001$).

General discussion and conclusion

Across two preregistered studies, we tested whether observers infer numerical competence in a way that is consistent with a normative Bayesian account, and whether they additionally incorporate response-time information when available. In Study 1, participants predicted another person's performance from minimal evidence—a single success or failure on one mental arithmetic problem—and their judgments closely tracked objective conditional probabilities. Participants' behavior was well approximated by a Bayesian model optimally integrating new performance information with prior expectations. In Study 2, we extended the paradigm by providing participants with information about a time-limit constraint (Fast vs. Slow) along with the observed performance. Although the time-limit cue was (weakly) diagnostic of competence in the performance data, participants did not reliably use it when making judgments.

A first contribution of this work is to extend Bayesian accounts of competence inference from knowledge-based (e.g. Aboody et al., 2025; Mercier et al., 2026) to skill-based competence. Numerical competence differs from knowledge attribution in that performance depends not chiefly on stored facts but mainly on computation under time constraints. We found that participants have a well-calibrated estimation of the difficult and that they relied on a similar inferential logic to infer either knowledgeability or numerical ability. Moreover, accurate inference did not require high arithmetic ability: even the lowest-performing participants tracked difficulty and accurately predicted the conditional probability of solving a math problem.

Study 2 provides a boundary test for cue integration. Holding accuracy constant, the Fast (20 s) versus Slow (40 s) constraint carried some objective information about competence: among those who succeeded (or failed) on an observed item, performance under the faster time limit predicted a slightly higher probability of success on other items. However, participants' judgments were not influenced by the time-limit information. One interpretation is that observers prioritize robust cues (difficulty and accuracy) while treating speed as ambiguous or unreliable. Indeed, a binary label (within 20 s vs. within 40 s) collapses a wide range of possible response times and may therefore provide only a weak signal. This pattern appears to differ from developmental work suggesting that children use speed to infer competence when choosing between informants (Leonard et al., 2019). A key difference is that in Leonard et al. (2019) children observe differences in agents' efficiency (without an exogenous time-limit), and that work also documents boundary conditions: children's competence inferences weaken when the relevant cues are harder to discriminate or when speed conflicts with outcome. From this perspective, our results suggest that adults may require clearer, more directly attributable speed information to treat it as diagnostic in the presence of strong accuracy and difficulty cues.

A complementary interpretation is that there was limited margin for time information to improve judgments. Participants' predictions already tracked true conditional probabilities closely in Study 2, leaving only marginal gains for an additional cue with a small objective effect. Combined with binary predictions, modest shifts in inferred competence may rarely cross the threshold required to change a yes/no response, making cue use difficult to detect. Future work can adjudicate these accounts by eliciting graded probability judgments, presenting continuous response times (rather than a time-limit label), increasing the magnitude or reliability of speed diagnosticity, and extending the computational model to include additional latent factors such as caution or vigilance. More broadly, these findings suggest that competence inference can approximate an ideal Bayesian model across domains, while also highlighting principled limits on spontaneous integration of secondary cues when their incremental diagnostic value is small or context-dependent.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. We acknowledge financial support from the Agence Nationale de la Recherche, France (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 reference ANR-10-IDEX-0001.

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

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

Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023, May). Thinking about Thinking as Rational Computation. https://doi.org/10.31234/osf.io/e65p3

Cheng, D., Shi, K., Wang, N., Miao, X., & Zhou, X. (2021). Examining the Differential Role of General and Specific Processing Speed in Predicting Mathematical Achievement in Junior High School. *Journal of Intelligence*, *10*(1), 1. https://doi.org/10.3390/jintelligence10010001

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

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. OUP USA.

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

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*(2-3), 108–119. https://doi.org/10.1016/j.intell.2011.02.001

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

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. *Psychological Bulletin*, *144*(11), 1200–1227. https://doi.org/10.1037/bul0000164

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*.

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

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*(1), 91–98. https://doi.org/10.1016/S0022-1031(03)00065-9

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*.

Mercier, M., Morin, O., Mercier, H., & Quillien, T. (2026). Who knows what? Bayesian competence inference guides knowledge attribution and information search. *Cognition*, *273*(106533). https://doi.org/10.1016/j.cognition.2026.106533

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

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Roy, E., Guillaume, M., Van Rinsveld, A., Project iLead Consortium, Anguera, J. A., Bunge, S. A., Gazzaley, A., Hoeft, F., Mishra, J., Rosenberg-Lee, M., Uncapher, M. R., & McCandliss, B. D. (2025). Tablet-based arithmetic fluency assessment reveals developments in math cognition and math achievement from childhood to adolescence. *npj Science of Learning*, *10*(1), 19. https://doi.org/10.1038/s41539-025-00314-5

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.