nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01057-0

A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers

Rachel A. Jansen[1]✉, Anna N. Rafferty[2] and Thomas L. Griffiths[3]

Evaluating one's own performance on a task, typically known as 'self-assessment', is perceived as a fundamental skill, but people appear poorly calibrated to their abilities. Studies seem to show poorer calibration for low performers than for high performers, which could indicate worse metacognitive ability among low performers relative to others (the Dunning–Kruger effect). By developing a rational model of self-assessment, we show that such an effect could be produced by two psychological mechanisms, in either isolation or conjunction: influence of prior beliefs about ability or a relation between performance and skill at determining correctness on each problem. To disentangle these explanations, we conducted a large-scale replication of a seminal paper with approximately 4,000 participants in each of two studies. Comparing the predictions of two variants of our rational model provides support for low performers being less able to estimate whether they are correct in the domains of grammar and logical reasoning.

In copious work studying adult metacognition, participants appear to be miscalibrated in their ability to judge their own performance across a large variety of domains[1–3]. Although there are age-related improvements in metacognitive abilities whereby very young children overestimate their competence a great deal more than adults[4], as well as differences by domain[5,6], on most tasks, researchers find that accuracy is low when making judgments about one's performance, when estimating either one's score or standing relative to others[3]. There have been studies of specific domains such as weather forecasting[5] and particular ways of eliciting judgments[7] where participants do show much better calibration to their own abilities, but, in most settings, accuracy is typically limited. In an influential paper, Kruger and Dunning conducted a series of studies that suggested that poorer performers tended to be less well calibrated in their ability to judge their performance after completing a task than higher performers[8]. They construed poor perceived performance by the lowest-scoring individuals as a metacognitive deficit: the worst performers lacked the skills needed to correctly do the task and also to judge their performance on the task. Commonly known as the Dunning–Kruger effect, this theory continues to be featured regularly in the media[9], particularly for the purpose of rationalizing others' seemingly irrational behavior (for example, anti-vaxxers[10] and government officials[11]).

Although discussions of the overconfidence of poor performers have focused on the idea that these people are less sensitive to their own errors, thinking about self-assessment from the perspective of a rational agent potentially offers a different account. If we imagine individuals as naive statisticians analyzing their own behavior, the rational Bayesian solution is to combine the evidence from experience with one's prior beliefs. If those prior beliefs are that one will perform relatively well, this should lead to poor performers overestimating their ability and good performers underestimating their ability to at least some extent[12].

This alternative explanation engages with a different point than previous controversy over the Dunning–Kruger effect. Krueger and Mueller[13] argued that the effect could be a statistical artifact of regression to the mean or a general poor calibration among participants that leads most estimates to converge around the mean, paired with a 'better-than-average' effect. Kruger and Dunning[14] responded that regression to the mean did not adequately explain their results after a re-analysis of their original data. Through additional studies in more real-world settings and a meta-analysis, researchers[2] concluded that the Dunning–Kruger effect was the best interpretation of these new data. By correcting for potential measurement error, they compared this hypothesis with other prominent accounts, including regression to the mean[13] and a task difficulty account in which perceived difficulty of the task produces a general trend of overestimation or underestimation[15]. Although this debate has helped establish that the results often explained by the Dunning–Kruger effect are not due simply to a statistical artifact of regression to the mean produced by the data, there remains the possibility that the internal regression to the (prior) mean produced by rational Bayesian inference—a psychological rather than statistical explanation for the data—is driving the effect.

To tease apart these two possible psychological explanations for the Dunning–Kruger effect, we developed a mathematical framework for specifying rational models of self-assessment that formalizes the nuances of the competing psychological theories. Using this framework, we show that the two different theories predict different forms for the relation between performance and self-assessment. We next ran a pair of large-scale replications to more precisely identify the actual form of this relation, which has typically been measured relatively coarsely in previous research. This more fine-grained picture allows us to identify which psychological explanation best accounts for people's errors in self-assessment.

Model

There have been a variety of computational models of self-assessment, some of which have been focused on alternative explanations for the Dunning–Kruger effect, whereas others have focused on making

[1]Department of Psychology, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Computer Science, Carleton College, Northfield, MN, USA. [3]Department of Psychology, Princeton University, Princeton, NJ, USA. ✉e-mail: racheljansen@berkeley.edu

sense of a variety of different methods and types of self-assessment. In one general model of self-assessment[16], the authors took into account confidence and error detection to unify various methods of measuring self-assessment. This model's parameters represented a participant's 'sensitivity' and 'bias', where sensitivity is their ability to discriminate between correct and incorrect performance, and bias is a penchant toward high confidence ratings. This model is based on a signal detection approach and aims to develop a unified computational account of metacognition that accounts for both confidence and error detection. Formally, it asserts second-order computation as the means by which humans judge their own confidence and assumes that confidence is determined not just on the basis of a person's actions but along with knowledge of the covariance between decisions and metacognitive states. Healy and Moore[12] developed a formal model to contrast expected outcomes on the basis of the type of self-assessment measured, specifically comparing overestimation of score and overplacement in comparison with others.

Others have used additional analytical methods to contest the existence of the Dunning–Kruger effect. In one instance, researchers sought to demonstrate how measurement error could account for most of the Dunning–Kruger effect by formalizing this phenomenon in terms of true skill and overconfidence[17]. In their sample, they found a relation between estimated and true ability, but it was much weaker than expected. Another group[18] set out to demonstrate a problem with biased subject pools in papers demonstrating the Dunning–Kruger effect[8] but did not compare their model with actual data. Schlösser et al.[19] refuted this account by showing that its statistical assumptions were inconsistent with data from the original paper by Kruger and Dunning[8]. In contrast to these previous efforts, we constructed different versions of a computational model of self-assessment that instantiate the theories we aim to distinguish between. Fitting these competing models to data and performing a model comparison allows us to evaluate which theory provides a better explanation of the data. In doing so, we are able to explore subtle differences in the assumed mechanisms behind the effect: specifically, whether we need to postulate a metacognitive deficit among underperformers (as originally suggested by Kruger and Dunning) or whether assumptions about priors (as instantiated in previous Bayesian models[12]) are sufficient.

We specifically model absolute self-assessment (where participants estimate their total score after an assessment) and introduce parameters to adjust perceived prior ability in a domain, difficulty of the assessment and competence at accurately concluding whether an individual problem was solved correctly or not. These factors are similar to those identified by other researchers as contributing to poor absolute self-assessment. For instance, poor self-assessment has been linked to lack of insight into one's errors[2], similar to the idea of 'sensitivity' in previous models[16]. In addition, a claim has been made that a person's 'self-concept' forms the foundation for participants' judgments about their performance[20]. This is akin to the 'bias' parameter in previous models[16]. A separate research group[21] additionally labeled the relevant components of self-assessment as 'bias' and 'discrimination'. Here, we identify a computational approach that incorporates similar parameters, one corresponding to perceived ability in a domain and another to discrimination ability. We additionally integrate a difficulty parameter, motivated particularly by results indicating that self-assessment ability is also dependent on item difficulty[6,15], although we leave manipulating this parameter for future studies. We designed two versions of our model, one that makes predictions on the basis of simple Bayesian inference and another that links skill with metacognitive ability to mathematically describe a Dunning–Kruger effect. We compare these versions to explore potential reasons behind inaccurate self-assessment by teasing apart specific features of self-assessment, which we can use to evaluate the conclusions drawn by earlier research into the matter.

In our basic model of self-assessment, we assume that people's inferences about their ability are based on three factors: 1) beliefs about the correctness of individual responses, 2) beliefs about their own ability and 3) the difficulty of the task they are performing. We then conduct a rational analysis, in the spirit of Anderson[22], considering how a rational agent should solve the problem of estimating their ability conditioned on the observed data and their prior beliefs. The notion of rational behavior used here is slightly different from classical rationality, as we condition on people's beliefs without asking whether those beliefs are well calibrated to the environment. This allows us to explain behavior in terms of these beliefs, in accordance with other rational models of cognition[23]. The rational solution to estimating one's ability is now to use Bayesian inference, modeling someone's posterior beliefs about their ability following an assessment as a function of their beliefs about their ability before the assessment and about the difficulty of that assessment (the priors) and beliefs about their performance on each individual problem (the likelihood).

Because in this model the participant is making inferences on the basis of their own beliefs, the likelihood is person p's belief about correctness on item i, where they believe they are either correct ($X_{p,i}=1$) or incorrect ($X_{p,i}=0$). The likelihood is dependent on the difficulty of an item i (β_i) and the perceived ability of person p (θ_p). If a person has perfect knowledge of whether they answered correctly, then the one-parameter item response theory (IRT) model, known as a Rasch model[24], is a reasonable way for people to make inferences about their own ability. Rasch models are widely used for estimating student ability in the psychometrics literature and can also be seen as a simple logistic regression model. This track record of previous use and the simplicity of the model led us to favor an approach based on an IRT function:

$$P(X_{p,i}=1|\theta_p,\beta_i)=\frac{1}{1+\mathrm{e}^{-(\theta_p-\beta_i)}} \tag{1}$$

Note that, given the perceived ability and difficulty parameters, the probability of believing that one gave an incorrect response is one minus the probability of believing that one gave a correct response: $P(X_{p,i}=0|\theta_p,\beta_i)=1-P(X_{p,i}=1|\theta_p,\beta_i)=\frac{1}{1+\mathrm{e}^{(\theta_p-\beta_i)}}$.

Equation (1) assumes that people have perfect knowledge about whether they have answered a problem correctly, but, in reality, people might not know the correctness of each of their responses with certainty. To account for this fact, we combine the Rasch model with uncertainty about whether the person is correct via an error parameter ϵ. This term corresponds to the probability that a person is erroneous in their belief about whether they have correctly solved a problem. This additional parameter is similar to the 'sensitivity' parameter described above[16] or the idea of poor error detection[2]. Thus, the likelihood equation where we include an error parameter to adjust for uncertainty in people's beliefs about their correctness becomes

$$P(X_{p,i}=1|\theta_p,\beta_i;\epsilon)=(1-\epsilon)\cdot\frac{1}{1+\mathrm{e}^{-(\theta_p-\beta_i)}}+\epsilon\cdot\frac{1}{1+\mathrm{e}^{(\theta_p-\beta_i)}}. \tag{2}$$

The probability of the person believing they have responded correctly is then the probability of answering correctly and recognizing that one is correct plus the probability of answering incorrectly but erroneously believing one is correct.

On a multiple-choice assessment, participants might answer some questions correctly by guessing, even if they are unaware of which answer is correct or whether they are guessing correctly. Because people are likely to be aware that they are guessing some answers correctly at random, we add a guessing parameter, g, using

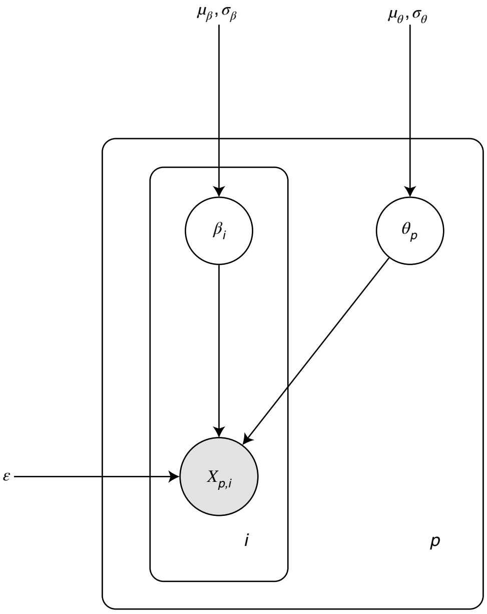


Fig. 1 | Graphical representation of the model. Each observed item $X_{p,i}$ is influenced by latent variables β_i (difficulty of problem i) and θ_p (perceived ability of person p) as well as a constant ϵ (ability to determine correctness), where the observed $X_{p,i}$ refers to a person's beliefs about correctness on an item. Difficulty β_i is drawn from a normal distribution with mean μ_β and standard deviation σ_β, and ability θ_p is drawn from a normal distribution with mean μ_θ and standard deviation σ_θ.

a simple variant of the IRT model that is commonly applied to multiple-choice assessments to account for this additional source of uncertainty:

$$P(X_{p,i}=1|\theta_p,\beta_i)=g+\frac{1-g}{1+\mathrm{e}^{-(\theta_p-\beta_i)}} \tag{3}$$

where $g=\frac{1}{N}$ and N represents the number of possible answers. This can also be modified to incorporate imperfect knowledge as in Equation (2), yielding

$$\begin{aligned} P(X_{p,i}=1|\theta_p,\beta_i;\epsilon) = \quad & (1-\epsilon)\cdot\left(g+\tfrac{1-g}{1+\mathrm{e}^{-(\theta_p-\beta_i)}}\right) \\ & +\epsilon\cdot\left(1-\left(g+\tfrac{1-g}{1+\mathrm{e}^{-(\theta_p-\beta_i)}}\right)\right) \end{aligned} \tag{4}$$

The priors are defined over the difficulty of an item i (β_i) and the perceived ability of person p (θ_p). Here, we assume the priors are normally distributed, although the model can use any prior distribution, which would allow for making more complex predictions. Varying the skew of the prior distribution over perceived ability θ_p, for example, would capture differing interpretations of successes and failures, such as learners being more likely to attribute a failure to a lack of ability rather than the task being difficult or vice versa.

A graphical model depicting the dependencies among the variables is shown in Fig. 1. To model people's posterior beliefs about their ability after performing a task given both their prior beliefs and the accuracy of their judgments of correctness, we insert the likelihood and priors into Bayes' rule as follows:

$$P(\theta_p,\beta_i|X_{p,i}=1)\propto P(X_{p,i}=1|\theta_p,\beta_i;\epsilon)\cdot P(\theta_p)\cdot P(\beta_i) \tag{5}$$

To calculate beliefs about performance from this posterior distribution, we first marginalize over β_i, which represents someone's posterior beliefs about the difficulty of the current assessment. We then transform the distribution to one about performance expectations rather than ability by calculating the probability of a correct response on an average-difficulty question ($\beta_i=0$) using Equation (1) for each point in the distribution. This transformation enables direct comparison of human data with our model's output.

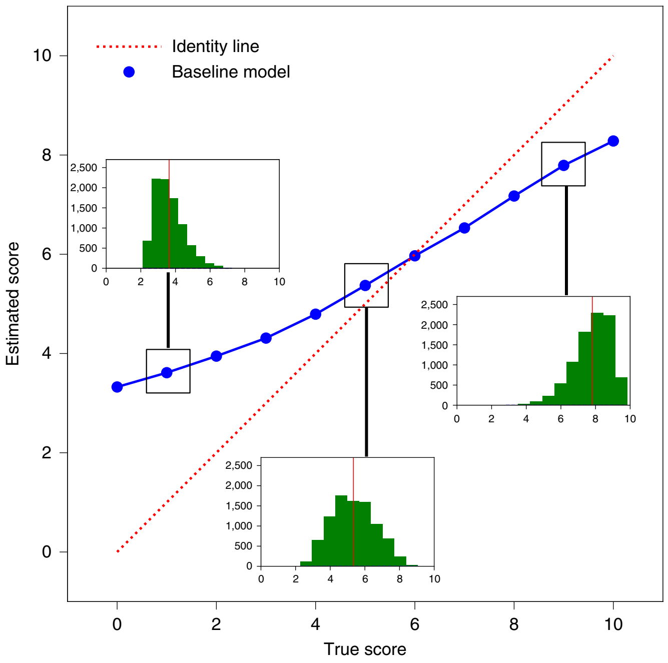


Fig. 2 | Model predictions in a toy example where participants solve ten problems in the baseline model (μ_θ, $\mu_\beta=0$, σ_θ, $\sigma_\beta=1$ and $\epsilon=0$). Each point shows the average of the posterior distribution on θ, corresponding to estimated ability. Overlaid are histograms of MCMC estimates of the posterior distribution on θ for three scores. The red dotted identity line represents completely accurate estimation, where a participant's estimated score is equal to their true score. The guessing parameter g is equal to 0.2, which represents a task with five multiple-choice solutions per question.

Bayesian inference. Here we demonstrate through simulations that this basic model predicts that people will not be fully accurate in their self-assessments and that poor performers will tend to overestimate whereas high-scoring individuals underestimate, consistent with Bayesian inference shifting estimates toward the mean of the prior. Changing the mean of the prior over the ability parameter θ_p adjusts the degree to which these patterns emerge. We begin by exploring a version of the model in which people are always accurate in their judgments of correctness on individual problems, which is instantiated by setting $\epsilon=0$, and then show what happens when ϵ is increased. In this model, we assume both constant mean of the prior distribution over θ_p and constant ϵ across individuals. We set the guessing parameter $g=0.2$, which assumes that there are five possible choices offered for each question.

We first consider what patterns in self-assessment occur when we assume people make perfectly accurate assumptions about their performance on each problem, by setting ϵ equal to 0. For these simulations, we assume θ_p and each β_i are distributed normally such that their means μ_θ and μ_β are 0 and their standard deviations σ_θ and σ_β are 1. As shown in Fig. 2, simulated participants from a toy example who perform on the low end tend to overestimate their performance, whereas the highest performers slightly underestimate their score, demonstrating a pattern of results similar to those found by the original authors[8]. To compute estimated performance given true score, we use Markov chain Monte Carlo (MCMC) methods[25] to calculate a posterior distribution over beliefs about ability

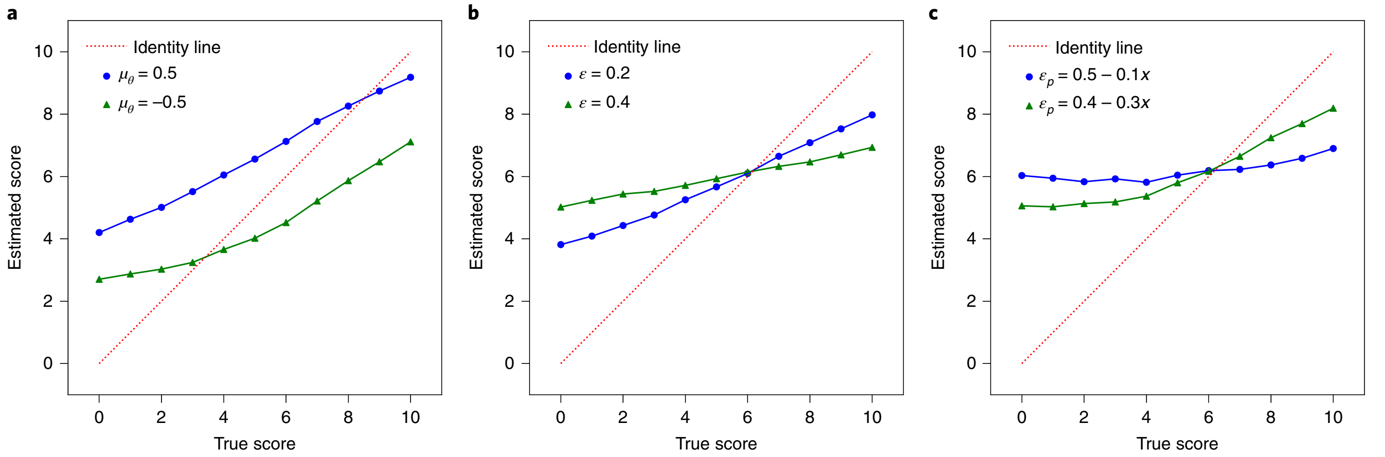


Fig. 3 | Model predictions in a toy example. a–c, Participants solve ten problems when the mean on ability (θ_p) is adjusted ($\mu_\theta = 0.5$ or -0.5) (**a**), or the parameter ϵ is adjusted ($\epsilon = 0.2$ or 0.4) (**b**), demonstrating the Dunning–Kruger effect (**c**). Note that x in the equations in **c** refers to normalized score (score divided by maximum score). Again, we set $g = 0.2$.

(Equation (5)) for each true score and then transform the result into a distribution over beliefs about performance. To obtain the estimated total score from this distribution, we scale the probability of being correct on a single problem by the maximum score (here, ten) and compute the expectation.

We run MCMC with 10,000 iterations and remove the first 1,000 as a 'burn-in', following standard practice, before taking the mean predicted score estimate. These baseline model predictions demonstrate that self-assessment ability need not be dependent on people's actual ability to obtain this pattern. Our rational model makes it straightforward to evaluate the consequences of changing people's prior expectations about their ability (the prior on ability parameter, θ_p) or their skill at recognizing whether they are correct on each problem (ϵ). Changing these aspects of the model has direct consequences for the form of the function relating estimated ability to true score.

To retrieve patterns of results even more similar to those found in previous research on self-assessment, we adjust the model parameters. Varying the prior via the mean, μ_θ, of the ability parameter, θ_p, changes the overall assessment of ability. As shown in Fig. 3a, when the mean on θ_p is high ($\mu_\theta = 0.5$), reflecting optimism about ability, there is much more overestimation by all simulated learners. But when the mean is lowered ($\mu_\theta = -0.5$), reflecting pessimism, we see the manifestation of the opposite pattern: except for all but the lowest performers, the model predicts underestimation rather than overestimation. The pattern of less underconfidence of high performers compared to overconfidence of low performers, the hallmark of the Dunning–Kruger effect, can be fit by this simple Bayesian inference model if we just increase the mean of the prior over ability θ_p.

Although changes to the prior affect the intercept of the line, changing ϵ affects its slope. As shown in Fig. 3b, as ϵ increases, the slope of the line decreases. In other words, as inferences about correctness become more similar to guessing randomly (which would be captured by $\epsilon = 0.5$), inferences about ability are predicted to become increasingly similar to one another, regardless of actual performance. Similar patterns of results are produced by manipulating the standard deviation of the mean on θ_p, σ_θ, which are detailed in the Supplementary Information. Both parameters affect the influence of performance on self-assessment. Although we will keep σ_θ fixed and vary ϵ, the conclusions we draw as a result of this apply to this broader capacity for updating beliefs about ability on the basis of performance rather than any specific parameter.

Performance-dependent estimation. The model described so far has assumed that everyone is equally adept at knowing whether their responses were correct or incorrect, consistent with explanations that assume similar metacognitive abilities on average regardless of true ability[13,15]. The idea that poor performers are 'metacognitively impaired' in comparison with high performers, as put forth by Kruger and Dunning[8], can be captured by extending the model so that, instead of an ϵ parameter that is identical across all participants, there is, instead, a separate ϵ_p associated with each person p that might differ across individuals in relation to their true ability (μ_θ remains constant across all individuals in this model). Namely, those who perform poorly will guess their performance on a single problem less accurately than those who perform well: ϵ_p for lower performers will be larger than for higher performers.

One way to make ϵ_p dependent on person p's ability is to use a simple linear function such that ϵ_p varies linearly with score, which serves as our closest approximation of true ability:

$$\epsilon_p = \epsilon_0 - \alpha \cdot \frac{\sum_i x_i}{n}, \qquad (6)$$

with slope $-\alpha$, intercept ϵ_0 and maximum achievable score n. Thus, $\frac{\sum_i x_i}{n}$ represents someone's scaled score. In the example in Fig. 3c, we vary ϵ_p gradually according to Equation (6) with $\epsilon_0 = 0.5$, $\alpha = 0.1$ and then with $\epsilon_0 = 0.4$, $\alpha = 0.3$. In these examples, the worst performers have an ϵ_p of 0.5 (at chance) or 0.4, and the highest performers have a slightly lower ϵ_p (0.4 in the first example and 0.1 in the second), meaning that they are more accurate in their beliefs about correctness. This produces greater overestimation at lower true scores.

To evaluate whether performance-dependent estimation—the explanation that Kruger and Dunning gave for their results[8]—actually occurs, we can compare how well two versions of the model that capture competing hypotheses about self-assessment fit the human data. The two variants of our Bayesian model are 1) a Bayesian inference model (where ϵ is independent of true ability), which represents a simple explanation of the data, and 2) a performance-dependent estimation model (where ϵ_p is dependent on person p's score). It is worth noting that the two models differ most in their predictions about people's beliefs in the tails of the plots, so experimentally differentiating between theories will require recruiting sufficient numbers of participants with extreme scores.

Previous arguments in favor of the Dunning–Kruger effect have performed an analysis by grouping participants on the basis of quartile of performance[2,13]. Having our distinct models will help tease apart possible interpretations of the data, but the way the data have

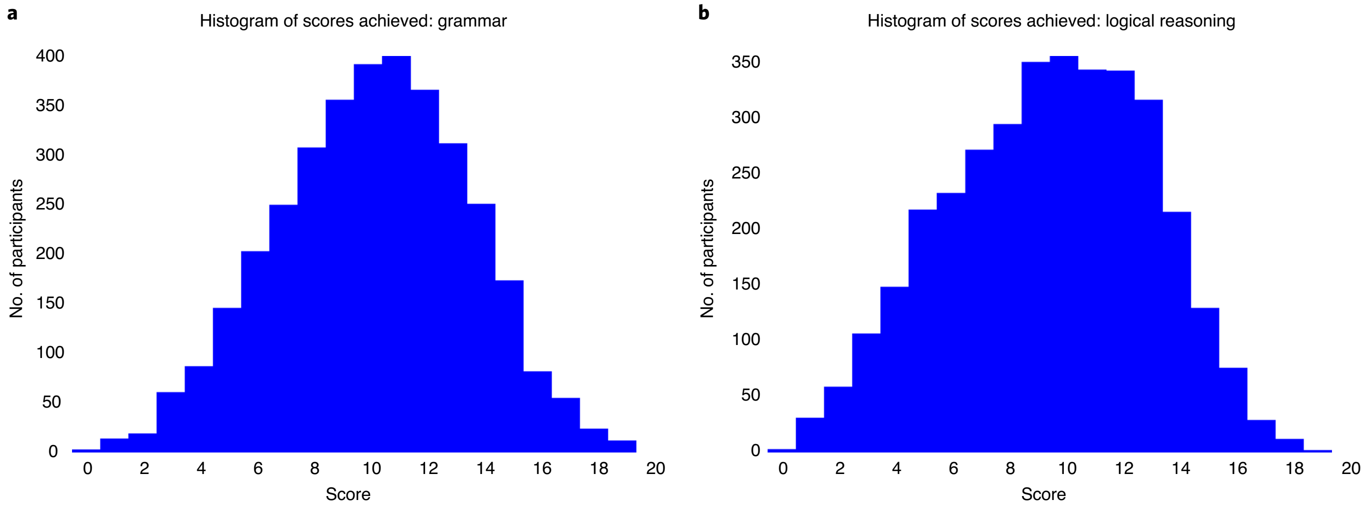


Fig. 4 | Histograms of results in the two studies. a,b, Scores achieved in the grammar study (**a**) and logical reasoning study (**b**).

been looked at previously is also not sufficiently high resolution to demonstrate differences between the model fits because it provides only four data points (one for each quartile) to compare with a model. Therefore, we compare each individual data point with the model rather than grouping the data by quartile of performance.

These considerations argue for conducting a large-scale replication of previous experiments on the Dunning–Kruger effect, which did not have sufficiently large samples to distinguish between the two explanations instantiated in our models. To address this, we conducted replications of two studies from Kruger and Dunning[8] in which participants solved a series of 20 multiple-choice questions about either grammar or logical reasoning and estimated their score after the assessment.

Results

Grammar study. Of the 3,515 participants who solved the grammar problems included in analyses (1,698 self-identifying men, 1,780 self-identifying women and the remainder other or unspecified; 2,560 White, 304 Black or African American, 246 Asian/Asian American, 215 Hispanic or Latino and the rest selecting multiple categories, other or unspecified), the mean completion time was 19.61 min. On average, participants scored 10.17 out of 20 (s.d. = 3.40) (see Fig. 4 for the distribution of the achieved scores), and the mean estimated score was 12.49 (s.d. = 3.91). In the original study by Kruger and Dunning, participants scored an average of 13.3 and estimated the number correct at an average of 15.2. We attribute this difference to the fact that their participants were undergraduate students, whereas participants in this study had a wide range of backgrounds. The original authors did not report standard deviations. The overconfidence of the lowest-scoring participants appeared substantial, as can be seen in Fig. 5. Participants made an average percentile estimate of 58.48 (s.d. = 20.57) and rated the task with a difficulty of 5.57 (s.d. = 2.26) for themselves and 6.16 (s.d. = 1.84) for other participants, both out of ten, where ten represents the highest difficulty.

To fit the model to the data, we compare model predictions to participants' estimates of their scores relative to their true score. Grouping the self-assessments by true score instead of by quartiles, as done in previous research, shows substantially more variability in performance. Comparing the data with the grid search of model estimates, the best-fit Bayesian inference model (where estimation ability is independent of score) was parametrized by $\epsilon = 0.4$ and $\mu_\theta = 0.05$ (s.s.e. = 49,591.58), whereas the performance-dependent estimation model (the model that instantiates the Dunning–Kruger effect) was parametrized by $\epsilon_0 = 0.45$, $\mu_\theta = 0.05$ and $\alpha = 0.1$ (s.s.e. = 49,073.04). The best-fit models are presented in Fig. 4a alongside the associated data, grouped by actual score.

To compare these competing models, as described in our analysis pipeline, we calculated the Bayesian information criterion (BIC) for each model. The BIC for the Bayesian inference model with constant ϵ (BIC 19,303.07) was higher than that of the model with ϵ dependent on score (BIC 19,274.29). Because these are nested models such that the performance-dependent model contains one more parameter than the Bayesian inference model, we performed a likelihood ratio test that is equivalent to a χ^2 test with one degree of freedom. This yielded $\chi^2(1) = 36.95, P < 0.001$ and $\phi = 0.62$, which far exceeds the threshold of 3.84 required to be significant. Thus, we have strong evidence to prefer the more complex model in this case, which is the performance-dependent model. We additionally computed the log Bayes factor, the logarithm ratio of the marginal likelihood of the performance-dependent hypothesis to the likelihood of the Bayesian inference hypothesis, and obtained a value of 16.14, which is another indication of strong evidence for the performance-dependent estimation hypothesis. Given that the parameters of our model correspond to the intercept (μ_θ), slope (ϵ_0) and curvature (α) of the data, we additionally fit linear and quadratic models to the data, finding that the quadratic model provided a better fit compared with the linear model ($F_1 = 34.25$, $P < 0.001$). When no exclusions are applied (as described in the Methods section), and data from all participants are included, results are substantially the same as those presented here.

We show in Fig. 5b a fit of the data to the model by quartile of performance, as done in previous work. Specifically, we group participants in quartiles on the basis of their scores and plot their average self-assessment judgments and overlay average model values for the best-fit models. We see clearly in this depiction that quartiles do not show the clear distinction between the two models that can be seen in the alternative plot.

Logical reasoning study. Of the 3,543 participants included in analyses who solved the logical reasoning problems (1,778 self-identifying women and 1,731 self-identifying men; 2,553 White, 350 Black or African American, 246 Asian/Asian American and 196 Hispanic or Latino), the average completion time was 23.48 min. The mean score was 9.45 out of 20 (s.d. = 3.59), and the mean estimated score was 10.86 (s.d. = 4.05). In the original study by Kruger and Dunning, participants scored an average of 12.9 and estimated an average of 13.3. As in the grammar study, we observed considerable overconfidence by the worst performers (Fig. 6). Participants

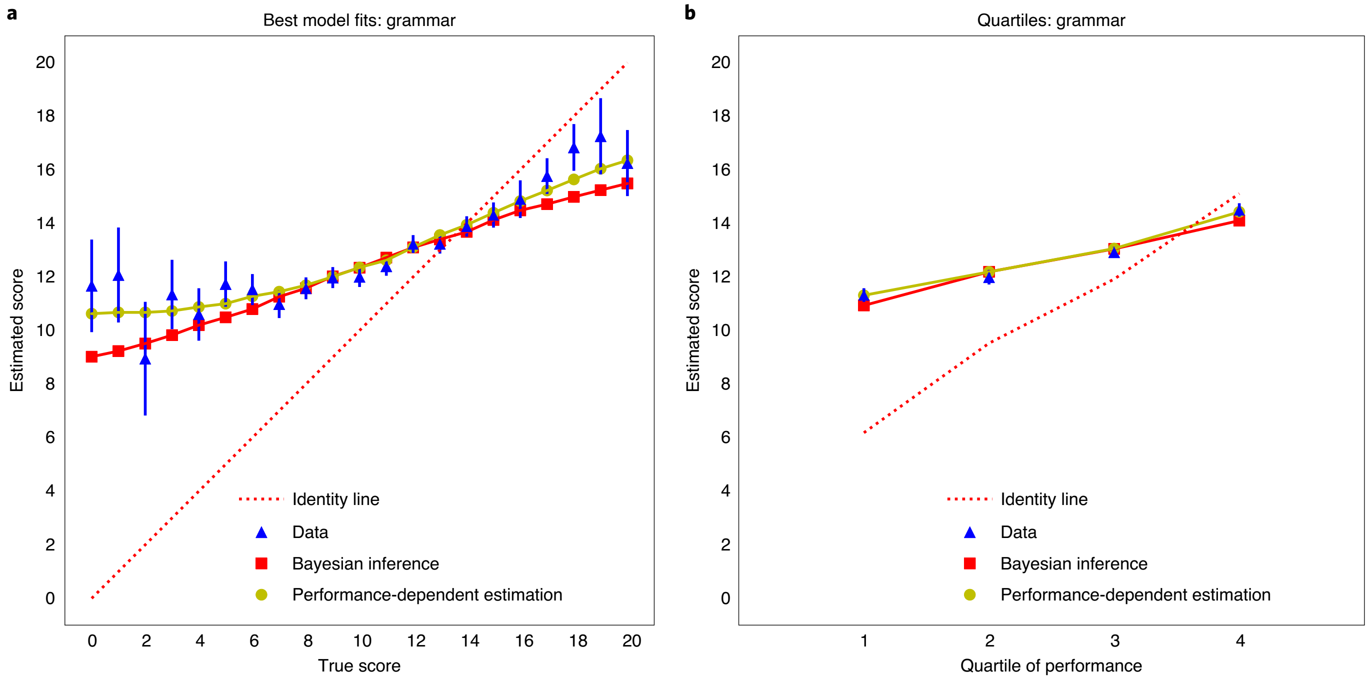


Fig. 5 | Findings of the grammar study. a,**b**, Results and best fitting models for displayed by score (**a**) and quartile of performance (**b**). The model where estimation accuracy is independent of score, or the Bayesian inference model, is parameterized by $\epsilon = 0.4$ and $\mu_\theta = 0.05$, whereas the performance-dependent model is parameterized by $\epsilon_0 = 0.45$, $\mu_\theta = 0.05$ and $\alpha = 0.1$. Error bars represent 95% confidence intervals.

made an average percentile estimate of 52.44 (s.d. = 20.84) and rated the task with a difficulty of 6.78 out of 10 (s.d. = 2.04) for themselves and 6.92 (s.d. = 1.79) for other participants.

Comparing the data with the grid search of model estimates, the best-fit Bayesian inference model was parameterized by $\epsilon = 0.45$ and $\mu_\theta = -0.1$ (s.s.e. = 55,801.41), whereas the performance-dependent estimation model was parameterized by $\epsilon_0 = 0.5$, $\mu_\theta = -0.15$ and $\alpha = 0.15$ (s.s.e. = 54,912.32), as shown in Fig. 6a. A quadratic model again was a better fit to the logical reasoning data as opposed to a linear model ($F_1 = 56.87$, $P < 0.001$).

The BIC for the model with constant ϵ (BIC 19,846.55) was again higher than that of the model with ϵ dependent on score (BIC 19,797.82). A likelihood ratio test comparing the models was significant ($\chi^2_1 = 56.91, P < 0.001$). Just as for the grammar study, when no exclusions are applied, results are substantially the same. A log Bayes factor calculation yielded a value of 26.15. Thus, we again have sufficient evidence to prefer the more complex performance-dependent estimation model over the Bayesian inference model.

Discussion

The observation that poor performers overestimate their ability could be given two possible psychological explanations: that it is a mere result of rational Bayesian estimation or that it reflects a genuine decreased sensitivity to errors among low performers. In this study, we formalized these competing accounts as mathematical models and ran large-scale replications to identify the form of the function relating self-assessment to performance. Although even with such large samples there is limited resolution in the tails of the score distributions (as fewer participants obtained the most extreme scores), the model assuming reduced sensitivity among low performers is a statistically better fit to the data than the model assuming simple Bayesian inference.

The rational model we developed gives us a framework for testing even more hypotheses beyond whether there is a relation between true ability and accuracy of perceived ability. Specifically, it offers a precise specification of the form of the function relating true ability to perceived ability, which allows for testing differences at task-specific and individual levels. This model enables future work that could include testing out a combination of priors over each parameter or identifying potential differences by domain, age or other demographic variables. We see this as a promising approach for gaining more insight into what underlies learners' metacognitive abilities. The flexibility of this model can also allow for fitting to different types of metacognitive data, including relative perceived performance (in this study, we focused on absolute estimates of performance) or incorporating new parameters to formalize other areas of research into metacognition, such as the bias blind spot[26].

There has been some debate as to whether self-assessment ability should be measured via one-off judgments (as in these studies) or by aggregating an individual's confidence judgments made before solving each problem[13]. Future directions for this research include implementing a version of the model that predicts a participant's confidence judgments over time to provide a more accurate representation of each individual's beliefs about their ability.

One possible limitation of our work is the use of Amazon's Mechanical Turk. Prior research has pointed out the limitations of this platform[27]. However, whereas participants from crowdsourcing services might not fully represent the population, they are far more representative than the undergraduate populations originally used in studies of the Dunning–Kruger effect—something that is reflected in differences between our results and those of Kruger and Dunning. In ongoing research, we are exploring the use of similar models with different populations, including specific age groups and clinical populations.

The large datasets that we collected through our experiments provide an impressively clear picture of the form of the relation between self-assessment and performance. Over the last few years, psychology has struggled with concerns about methodology and the replicability of prominent findings. We think that the way forward is not to merely try to replicate our past results but to combine the large sample sizes made possible by modern technology with advances in computational modelling to pursue new psychological

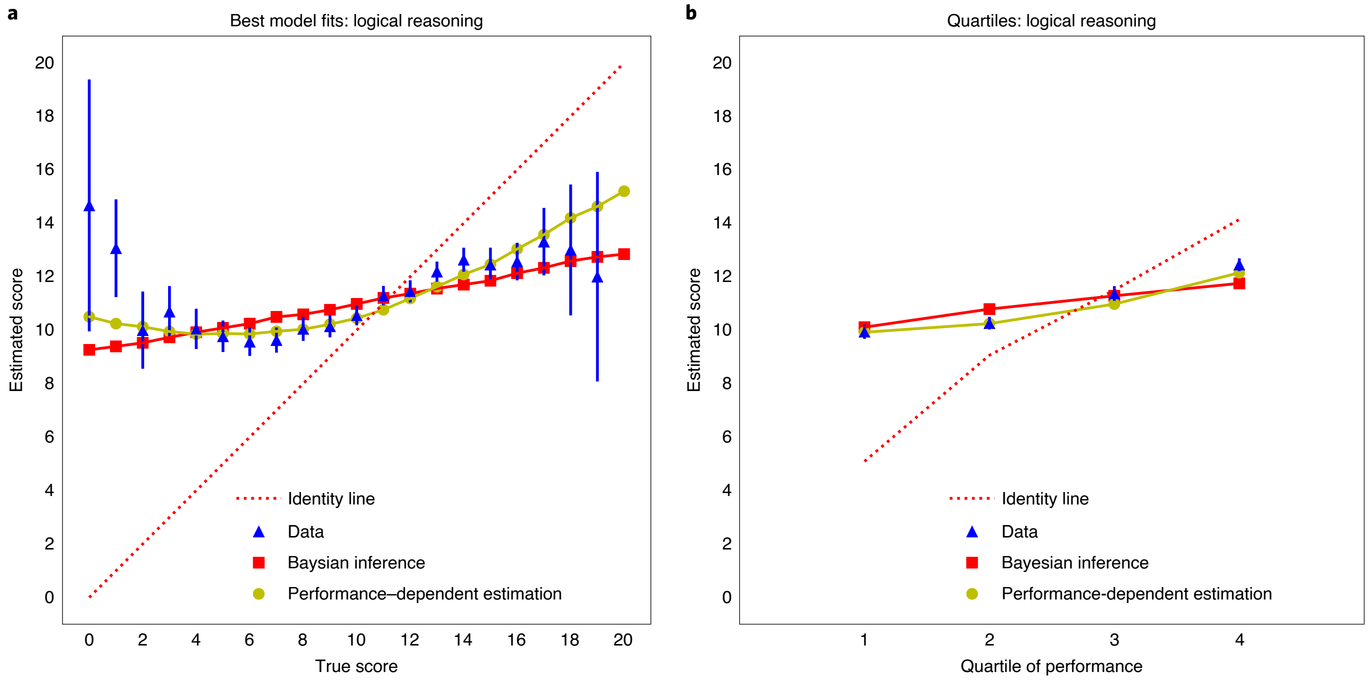


Fig. 6 | Findings of the logical reasoning study. a,**b**, Results and best fitting models displayed by true score (**a**) and quartile of performance (**b**). The simpler Bayesian inference model is parameterized by $\epsilon = 0.45$ and $\mu_\theta = -0.1$, whereas the performance-dependent model has $\epsilon_0 = 0.5$, $\mu_\theta = -0.15$ and $\alpha = 0.15$. Error bars represent 95% confidence intervals.

research that provides deeper insight into the phenomena behind those past results. As in the present case, the primary outcome of this research might be confirming the existence and interpretation of key phenomena, but providing definitive evidence for such phenomena—particularly those that enjoy the public profile of the Dunning–Kruger effect—is an important step toward re-establishing the validity of psychological research. We view our results as an example of what this approach can achieve, providing a high-resolution picture of the nature of human miscalibration.

Methods

This research complies with ethical obligations put forth by the University of California Berkeley Internal Review Board. Informed consent was obtained from all participants.

Participants. One criticism of Kruger and Dunning is their use of a convenience sample composed of college undergraduates at elite universities[18]. We allay this concern by recruiting participants from Amazon's Mechanical Turk (https://www.mturk.com/), where wider ranges of age and educational background are represented[28].

We selected our sample size of 4,000 participants per study by conducting a power analysis based on the effect sizes observed in preliminary studies of 250 participants each. To do so, we simulated increasing sample sizes and observed where the curves started to level out (see pre-registration (https://osf.io/k28je) from March 6, 2019). With 4,000 samples per bootstrap, 85.4% of the grammar data simulations and 81.6% of the logical reasoning simulations resulted in favor of the Dunning–Kruger effect. We declared a stopping rule (ending data collection once 4,000 responses were collected or after 10 d) as well as all exclusion criteria in our pre-registration. This power analysis was conducted with a previous version of the model that did not contain a guessing parameter. We re-ran our models to include this guessing parameter following a suggestion made by one of our reviewers.

Participants were paid $2 to participate in the grammar study or $3 for the logical reasoning study. We decided against awarding a bonus for more accurate self-assessment judgments because Kruger and Dunning also paid only a flat rate and because, in other work, adding incentives did not reduce the amount of inaccuracy[2,29]. Because sample sizes were so large, participants were allowed to participate in both studies, which were presented as separate human intelligence tasks on Mechanical Turk.

For the grammar study, there were 3,860 responses. Of these, 164 failed the instructional manipulation and so were excluded from analyses. Eighty spent under 5 min on the task. There were three internet protocol (IP) addresses used three times and 43 used twice, so these 95 responses were additionally excluded. Six responses had no associated IP address and were, thus, also excluded. This resulted in a total of 3,515 responses viable for analyses (1,698 self-identifying men and 1,780 self-identifying women; mean age 36.54 years, range 18–88 years).

The logical reasoning study brought in 3,901 complete responses. There were 154 failed attention checks, 77 participants who spent under 5 min and 55 repeated IPs (51 with two, 2 with three, 1 with four and 1 with six responses) and nine without an IP address. This left 3,543 responses for analyses (1,778 self-identifying women and 1,731 self-identifying men; mean age 36.59 years, range 18–81 years).

Materials. The closest approximation of the original 20 logical reasoning problems and 20 grammar questions from Kruger and Dunning[8] that we could obtain from the original authors were made into surveys on Qualtrics. We used their original materials to make a more compelling case for whether this effect exists.

These questions consist of multiple-choice items with five possible responses. In the logical reasoning test, participants were told 'You will be presented with brief passages or statements and will be required to evaluate their reasoning or determine what inferences you can logically draw from the passage. In each case, select the best answer choice, even though more than one choice may present a possible answer.' In the grammar task, participants read the following instructions: 'In each question, some part of each sentence is underlined; sometimes the whole sentence is underlined. Five choices for rephrasing the underlined part follow each sentence; one choice repeats the original, and the other four are different.'

Procedure. Before beginning each study, all participants read a set of instructions and were asked two content-based questions about the instructions. They were given two opportunities to answer these questions correctly and were excluded from analyses if they failed the instructional manipulation on both attempts, as indicated in our preregistration. These exclusions were intended to prevent the low-performing individuals in our analyses from being composed of inattentive participants. Both before and after problem solving, all participants rated their absolute performance ('how many of the 20 logical reasoning/grammar problems will/did you answer correctly?'), their relative performance as a percentile ranking out of 100 ('compared to other participants in this study, how well do you think you will do/did you do?'), the difficulty of the task for themselves and the difficulty for others (both on a scale from 0 to 10). On each task, the 20 questions were presented in a randomized order, as were the five multiple-choice solutions. All problems and self-assessment questions required a response for the participant to move ahead. The absolute performance ratings were displayed as a drop-down menu, the relative performance ratings as a sliding bar and the difficulty ratings as horizontal multiple-choice questions. At the conclusion of the study, participants were directed to a short demographics questionnaire where they optionally answered questions about their age, gender, race and educational background. All analyses in this paper are based solely on the absolute ratings of performance made after the test.

Model fitting. To fit the Bayesian inference and performance-dependent estimation models to the data, we compare model predictions with participants' estimates of their scores relative to their true score, because 'perceived performance' on the task is how we are operationalizing 'perceived ability'. To generate a set of simulations with varying parameter values, we performed a grid search over μ_θ and ϵ for the Bayesian inference model and these two parameters, along with α for the performance-dependent estimation model, such that values of $\mu_\theta \in [-1, 1]$, $\epsilon \in [0, 0.5]$ and $\alpha \in [0, 0.5]$ were considered. We do not consider values of ϵ greater than 0.5 or worse than chance because we assume people are not systematically biased to believe they are incorrect when they are correct. We took steps of 0.05 for each parameter, which resulted in a total of 41 considered values of μ_θ and 11 values each of ϵ and α, which produced 451 Bayesian inference model predictions and 2,706 performance-dependent estimation model predictions. Note that α cannot be lower than ϵ_0 in the performance-dependent estimation model, as this would result in negative values of ϵ_p, which is impossible because this is a probability, so there will not be $11 \times 11 \times 41 = 4{,}961$ performance-dependent estimation model predictions, as one might expect. Baseline values were used for the other parameters ($\sigma_\theta = \sigma_\beta = 1$; $\mu_\beta = 0$). We ran five MCMC chains of 10,000 iterations where we marginalized over item difficulties (β_i) to examine only perceived ability (removing the first 1,000 iterations each time for burn-in). We then took the mean of all remaining 9,000 sampled θ_p values to generate predictions for each pairing or triplet of parameters. To better estimate the posterior, we calculated the mean across all samples from all five chains to compare with the human data. We calculated all these values for the performance-dependent estimation model and used the values from the Bayesian inference model for the cases when $\alpha = 0$. We then converted each simulated ability parameter, θ, value generated by the models into a probability of a correct response using Equation (1). To transform this into an estimated total score, which is the data we have to compare with the model, we then multiply this probability by the maximum score (20 in our data).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The anonymized data that support the findings of this study are available on the Open Science Framework (https://osf.io/er9ms/).

Code availability

The Qualtrics code that generates the surveys is available in the same repository (https://osf.io/er9ms/) on the Open Science Framework. All code used for analyses and the model are available on GitHub (https://github.com/racheljansen/self-assessment).

Received: 4 May 2020; Accepted: 19 January 2021;
Published online: 25 February 2021

References

1. Dunning, D., Heath, C. & Suls, J. Flawed self-assessment: implications for health, education, and the workplace. *Psychol. Sci. Public Interest* **5**, 69–106 (2004).
2. Ehrlinger, J., Johnson, K., Banner, M., Dunning, D. & Kruger, J. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organ. Behav. Hum. Decis. Process.* **105**, 98–121 (2008).
3. Zell, E. & Krizan, Z. Do people have insight into their abilities? A metasynthesis. *Perspect. Psychol. Sci.* **9**, 111–125 (2014).
4. Bjorklund, D. F. & Green, B. L. The adaptive nature of cognitive immaturity. *Am. Psychol.* **47**, 46–54 (1992).
5. Tyszka, T. & Zielonka, P. Expert judgments: financial analysts versus weather forecasters. *J. Psychol. Financial Mark.* **3**, 152–160 (2002).
6. Jansen, R.A., Rafferty, A.N. and Griffiths, T.L. Algebra is not like trivia:evaluating self-assessment in an online math tutor. in *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, 2017).
7. Nelson, T. O. & Dunlosky, J. When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the 'delayed-JOL effect'. *Psychol. Sci.* **2**, 267–271 (1991).
8. Kruger, J. & Dunning, D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personal. Soc. Psychol.* **77**, 1121–1134 (1999).
9. Lopez, G. Why incompetent people often think they're actually the best. *Vox* https://www.vox.com/science-and-health/2017/11/18/16670576/dunning-kruger-effect-video (18 November 2017).
10. Andrews, R. This psychological effect explains why anti-vaxxers believe what they velieve. *IFLScience* http://www.iflscience.com/health-and-medicine/antivaxxers-suffer-from-a-wellknown-cognitive-effect-according-to-study/ (2018).
11. Purtill, C. This psychological quirk could explain why Trump's least experienced lawyer feels so confident. *Quartz* https://work.qz.com/1240245/the-dunning-kruger-effect-what-trumps-legal-team-and-the-russia-probe-have-to-do-with-it/ (29 March 2018).
12. Healy, P.J. & Moore, D.A. Bayesian overconfidence. *SSRN* https://doi.org/10.2139/ssrn.1001820 (2007).
13. Krueger, J. & Mueller, R. A. Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *J. Personal. Soc. Psychol.* **82**, 180–188 (2002).
14. Kruger, J. & Dunning, D. Unskilled and unaware-but why? A reply to Krueger and Mueller. *J. Personal. Soc. Psychol.* **82**, 189–192 (2002).
15. Burson, K., Larrick, R. P. & Klayman, J. Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *J. Personal. Soc. Psychol.* **90**, 60–77 (2006).
16. Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
17. Feld, J., Sauermann, J. & De Grip, A. Estimating the relationship between skill and overconfidence. *J. Behav. Exp. Econ.* **68**, 18–24 (2017).
18. Krajč, M. & Ortmann, A. Are the unskilled really that unaware? Alternative explanation. *J. Econ. Psychol.* **29**, 724–738 (2008).
19. Schlösser, T., Dunning, D., Johnson, K. L. & Kruger, J. How unaware are the unskilled? Empirical tests of the 'signal extraction' counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *J. Econ. Psychol.* **39**, 85–100 (2013).
20. Ehrlinger, J. & Dunning, D. How chronic self-views influence (and potentially mislead) estimates of performance. *J. Personal. Soc. Psychol.* **84**, 5–17 (2003).
21. Dunning, D. & Helzer, E. G. Beyond the correlation coefficient in studies of self-assessment accuracy: commentary on Zell & Krizan (2014). *Perspect. Psychol. Sci.* **9**, 126–130 (2014).
22. Anderson, J. R. *The Adaptive Character of Thought* (Earlbaum, 1990).
23. Oaksford, M. & Chater, N. A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* **101**, 608 (1994).
24. Embretson, S. E. & Reise, S. P. *Item Response Theory* (Psychology Press, 2013).
25. Gilks, W. R, Richardson, S. & Spiegelhalter, D. *Markov Chain Monte Carlo in Practice.* (Chapman and Hall/CRC, 1995).
26. Pronin, E., Lin, D. Y. & Ross, L. The bias blind spot: perceptions of bias in self versus others. *Personal. Soc. Psychol. Bull.* **28**, 369–381 (2002).
27. Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J. & Litman, L. Online panels in social science research: expanding sampling methods beyond Mechanical Turk. *Behav. Res. Methods* **51**, 2022–2038 (2019).
28. Mason, W. & Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* **44**, 1–23 (2012).
29. Sanchez, C. & Dunning, D. Overconfidence among beginners: is a little learning a dangerous thing? *J. Personal. Soc. Psychol.* **114**, 10 (2018).

Acknowledgements

The authors received no specific funding for this work.

Author contributions

R.A.J. designed the studies, collected data, developed the model and performed the simulations. A.N.R. and T.L.G. supervised the study design and model development. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-021-01057-0.

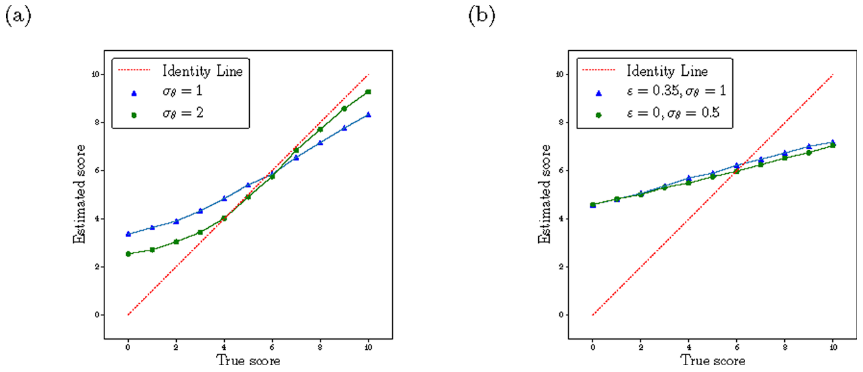
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01057-0.

Correspondence and requests for materials should be addressed to R.A.J.

Peer review information *Nature Human Behaviour* thanks Stephen Fleming, Sam Gilbert and Matthew Rhodes for their contribution to the peer review of this work. Primary Handling Editor: Marike Schiffer.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Interpreting σ_θ. Model predictions in a toy example where participants solve 10 problems (a) when the standard deviation on ability (σ_θ) is adjusted ($\sigma_\theta = 1$ or 2) and (b) when both this and the parameter ϵ are adjusted (σ_θ= 1, ϵ = 0.35 or ϵ = 0, σ_θ= 0.5) to reveal comparable results. In the main paper, we consider a single value for the standard deviation of the prior on ability (σ_θ). As shown in Fig. 1a, increasing the standard deviation of the prior implies more accurate estimation of scores, although some under and over estimation is still present. The pattern of Fig. 1a is similar to the pattern of predictions when changing ϵ. As shown in Fig. 1b, adjustments to either of these parameters can lead to very similar predictions for the relationship between true scores and estimated scores. This is not surprising given that both of these parameters represent uncertainty. Choosing to focus on fitting participants' values of ϵ allows us to capture variation in estimates of correctness on each question. On the other hand, if we were to focus on fitting participants' σ_θ values, we would be assuming variation in prior beliefs about ability. Given the framing of the Dunning–Kruger effect in terms of sensitivity to errors, we fixed σ_θ and focused on ϵ in our modeling approach. We have expressed our conclusions in terms consistent with variation in either ϵ or σ_θ, which affect the degree of updating of prior beliefs in light of evidence.

nature research

Corresponding author(s): Rachel Jansen

Last updated by author(s): Jan 8, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected on Amazon's Mechanical Turk through a Qualtrics survey.
Data analysis	All analyses and models were developed using Python code.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The anonymized data that support the findings of this study are available on the Open Science Framework (https://osf.io/er9ms/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative study
Research sample	Grammar study: 3,515 participants (1,698 men, 1,780 women, 2560 white, 304 Black or African American, 246 Asian/Asian American, 215 Hispanic or Latino), not fully representative. Logical reasoning study: 3,543 participants (1,778 women, 1,731 men, 2,553 white, 350 Black or African American, 246 Asian/Asian American, 196 Hispanic or Latino). Not representative.
Sampling strategy	Participants were allowed to enroll in the study through MTurk if they were based in the US. We aimed for a sample size of 4,000 for each study based on a power analysis conducted with pilot data, described in detail in the preregistration on OSF (https://osf.io.k28je/)
Data collection	Participants were recruited from Amazon's Mechanical Turk, so no experimenter was present when they were completing the task. The tasks were computer-based and all data was recorded through Qualtrics.
Timing	Data collection began on June 5, 2019 and concluded ten days later on June 14, 2019. Our preregistration stated that we would attempt to recruit 4,000 unique participants for each study (though MTurk workers could participate in both studies), but stop data collection after ten days.
Data exclusions	Grammar study: 164 failed instructional manipulation, 80 spent under 5 minutes, 95 duplicate IP addresses, 6 empty IP addresses Logical reasoning: 154 failed instructional manipulation, 77 spent under 5 minutes, 55 duplicate IP addresses, 9 empty IP addresses
Non-participation	Grammar study: 154 dropped out Logical reasoning study: 8 declined to participate and 112 dropped out.
Randomization	There were no experimental groups, but the order in which problems were presented to each participant was randomized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Participants were told from the start whether they were to answer logical reasoning or grammar questions, so there may be some self-selection bias in terms of participants wanting to solve something they knew they could succeed at. However, due to the broad range of scores, this did not seem to impact the results.
Ethics oversight	UC Berkeley Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.