Modeling the Dunning-Kruger Effect: A Rational Account of Inaccurate Self-Assessment

Rachel A. Jansen (racheljansen@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Anna N. Rafferty (arafferty@carleton.edu)
Department of Computer Science, Carleton College
Northfield, MN 55057 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Abstract

Self-assessment, or the evaluation of one's ability on a task, is widely perceived as a fundamental skill, yet in most studies, people are found to be poorly calibrated to their own abilities. Some results seem to show poorer calibration for low performers than for high performers. This effect has been explained in multiple ways: it could indicate worse metacognitive ability among the low performers (the "Dunning-Kruger" effect), or simply regression to the mean. To tease apart these explanations we develop a Bayesian model of self-assessment and evaluate its predictions in two experiments. Our results suggest that poor self-assessment is caused by the influence of prior beliefs and imperfect skill at determining whether a problem was solved correctly or not, and offer only weak support for of a relationship between metacognitive ability and performance.

Keywords: self-assessment; logical reasoning; metacognition; Bayesian modeling

Introduction

It has generally been found that people are miscalibrated in their ability to judge their own performance across many domains (e.g., Dunning, Heath, & Suls, 2004). Yet the potential causes of this phenomenon are not often agreed upon. In early work, Kruger and Dunning (1999) found that poorer performers tended to be less well-calibrated in their ability to judge their performance than higher performers. They interpreted 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. Krueger and Mueller (2002) disagreed with this conclusion, claiming the cause to be mere regression to the mean. Kruger and Dunning (2002) argued that this explanation still did not explain their original results.

Resolving this debate requires designing a formal account of self-assessment that makes it possible to evaluate the need for a dependence between ability and calibration. Taking a computational modeling approach makes it possible to directly compare theories about how performance and self assessment are related to experimental results. So far, computational models of self-evaluation have had other goals. Fleming and Daw (2017), for example, set up a model that takes into account confidence and error detection in order to unify different methods of measuring self-evaluation. Healy and Moore (2007) developed a formal model to contrast expected outcomes based on the type of self-assessments measured, specifically comparing overestimation of score and overplacement in comparison to others.

In this paper, we take the approach of modeling how a rational agent would self-assess as a starting point for modeling people. We specifically model *absolute* self-assessment (where participants guess 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. This allows us to tease apart the factors that contribute to self-assessments of ability.

In the next section, we begin by defining a rational model which offers multiple predictions about how self-assessment will play out under different circumstances and demonstrates a form of regression to the mean. Following this, we test the model predictions in one of the domains originally studied by Kruger and Dunning (1999): logical reasoning. We then present a more complex version of the model that allows ability and calibration to depend on one another and use a larger sample of data to compare between the simple and complex models. We find only weak evidence for the more complex model.

Modeling Self-Assessment

We assume that people's inferences about their ability are based on the correctness of their responses, their beliefs about their own ability, and the difficulty of the task they are performing. Because we are interested in modeling what a rational agent would do, it is natural to use a Bayesian formulation (e.g., Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), where we model someone's posterior beliefs about their performance following an assessment as a function of their beliefs about their ability before the assessment and the difficulty of that assessment (the *priors*) and their performance on each individual problem (the *likelihood*).

The *likelihood* is dependent on the response of person p to item i (X_{pi}), the difficulty of item i (β_i), and the perceived ability of person p (θ_p), where their response is either cor-

rect ($X_{pi} = 1$) or incorrect ($X_{pi} = 0$). Following the educational psychometrics literature, we assume this can be captured by a 1-parameter Item Response Theory (IRT) model known as a Rasch model (see Embretson and Reise (2013) for an overview):

$$P(X_{pi} = 1|\theta_p, \beta_i) = \frac{1}{1 + e^{-(\theta_p - \beta_i)}}. \quad (1)$$

Equation 1 assumes that people have perfect knowledge about whether they have answered a problem correctly, but in reality, learners may not know the correctness of each of their responses with certainty. To account for this fact, we include a parameter ε that corresponds to the probability of being mistaken about whether one is correct on a given problem. Thus the probability of believing one responded correctly is:

$$P(X_{pi} = 1|\theta_p, \beta_i, \varepsilon) = (1 - \varepsilon) \cdot \frac{1}{1 + e^{-(\theta_p - \beta_i)}} + \varepsilon \cdot \frac{1}{1 + e^{(\theta_p - \beta_i)}}. \quad (2)$$

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 Figure 1. To model people's beliefs about their abilities given both their prior beliefs and their judgment of correctness, we use Bayes' rule:

$$P(\theta_p, \beta_i|X_{pi} = 1) \propto P(X_{pi} = 1|\theta_p, \beta_i, \varepsilon) \cdot P(\theta_p) \cdot P(\beta_i). \quad (3)$$

Model Predictions

Using Markov chain Monte Carlo to sample from the posterior over perceived ability θ_p for different scores on an assessment, we retrieve a pattern of somewhat inaccurate estimation of performance (see Figure 2a). For each possible score out of ten, we sample perceived ability θ_p and a vector of β_is, and then integrate out β to obtain a marginal distribution over θ_p. We then convert each simulated ability parameter θ_p into the probability of a correct response on a new item j via Equation 1 (assuming $\beta_j = 0$). To obtain expected total score, we multiply by the maximum score (in this example, 10) and take the mean of all predicted values.[1] In Figure 2a, we set ε equal to zero (as though participants have perfect assumptions about their performance on each problem), and we assume θ_p and each β_i are distributed normally with μ_θ, μ_β equal to 0 and σ_θ, σ_β equal to 1. Simulated participants performing on the low end tend to overestimate their performance while the highest performers slightly underestimate their score, consistent with the pattern of results demonstrated in Kruger and Dunning (1999). These model predictions demonstrate that self-assessment ability need not be dependent on people's actual ability to get this pattern, consistent with the regression to the mean interpretation proposed by Krueger and Mueller (2002).

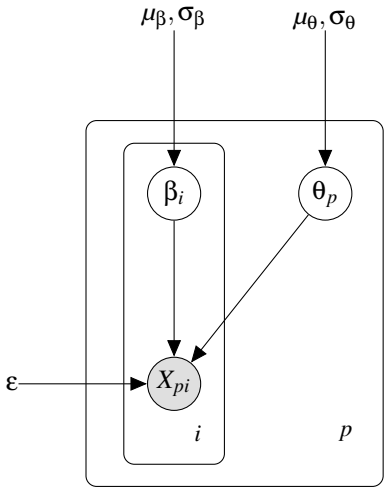


Figure 1: Graphical representation of the model: each observed response X_{pi} is influenced by latent variables β_i (drawn from the difficulty prior) and θ_p (perceived ability prior) as well as a constant ε (likelihood). β_i is drawn from a normal distribution with mean μ_β and standard deviation σ_β and θ_p is normal with mean μ_θ and standard deviation σ_θ.

Our rational model makes it straightforward to evaluate the consequences of changing people's prior expectations about their ability (the prior on 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.

Changes in the prior Varying the prior via the mean, μ_θ, of θ_p changes the overall assessment of ability. As shown in Figure 2b, when the mean on θ_p, the ability parameter, is high ($\mu_\theta = 0.5$), there is much more overestimation. But when the mean is lowered ($\mu_\theta = -0.5$), we see the manifestation of the opposite pattern: except for all but lowest performers, the model predicts under-estimation rather than over-estimation.[2]

Changes in the likelihood While changes to the prior affected the intercept of the line, changing ε affects the slope. As shown in Figure 2c, as ε increases, the slope of the line decreases. As inferences about correctness become more similar to guessing randomly (captured by ε = 0.5), inferences about ability are predicted to become more and more similar to one another regardless of actual performance.[3]

[1] We run this with 10,000 iterations and remove the first 1,000 for burn-in before taking the mean predicted score estimate.

[2] Similar patterns of results will be produced by manipulating the parameters over β_i, but we leave this for future work.

[3] Similar patterns of results will be produced by manipulating the

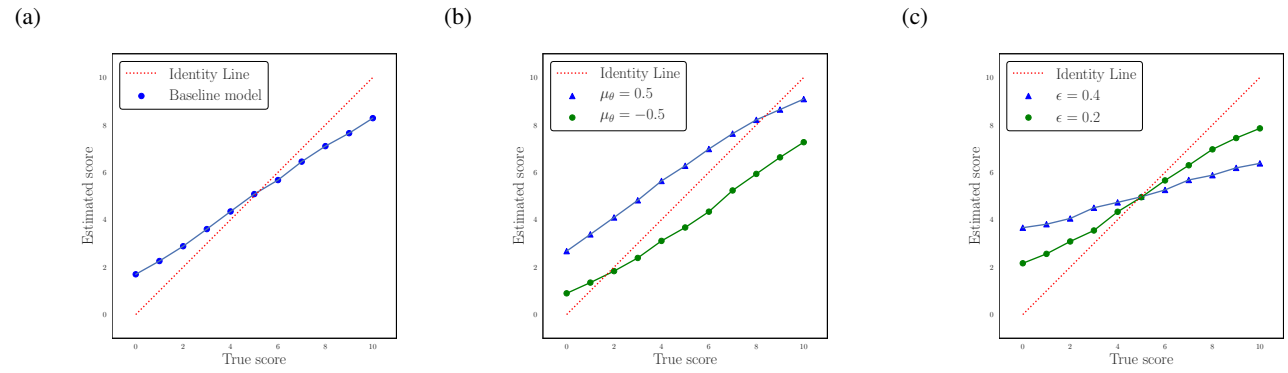


Figure 2: Model predictions for (a) the baseline model (μ_θ, $\mu_\beta = 0$, σ_θ, $\sigma_\beta = 1$, and $\varepsilon = 0$), (b) when the mean on ability (θ_p) is adjusted ($\mu_\theta = 0.5$ or -0.5), and (c) when the parameter ε is adjusted ($\varepsilon = 0.2$ or 0.4).

Experiment 1: Testing the Model

Our rational model predicts that people will inaccurately estimate their performance due to a combination of the effect of prior beliefs about their ability and imperfect skill at guessing their performance on each individual problem. To see how well models with differing parameters capture reality, we set up an experiment similar to those done by earlier researchers. One condition replicates the approach taken in previous work. In the other condition, participants were given immediate feedback after each problem. In our model, this can be interpreted as reducing the ε parameter – if people know whether they were right or wrong, ε should be effectively 0. As shown in Figure 2c, our model predicts this should attenuate the magnitude of the Dunning-Kruger effect.

Methods

Participants A total of 100 participants (50 per condition) were recruited on Amazon's Mechanical Turk (MTurk) and were each paid $3 for their time. To compensate for the fact that these are MTurk participants, some of whom were trying to complete the task as quickly as possible to obtain the payment, we eliminated those who spent under 10 minutes on the task. Additionally, since the population from Kruger and Dunning (1999) were undergraduates, we decided to exclude anyone who indicated that they had attended law school or taken the LSAT.

Procedure All participants completed 20 logical reasoning problems adapted from the 2007 LSAT (Law School Admissions Test).[4] This domain was selected because it was one of the domains initially studied by Kruger and Dunning (1999). All participants rated their absolute ability ("how many of the 20 logical reasoning problems will/did you answer correctly?"), their relative ability ("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. To analyze self-assessment, we focus on ratings of absolute ability. At the conclusion of the study, participants were directed to a short demographics questionnaire. The "no feedback" condition was a direct reproduction of the design used in previous studies. In the "feedback" condition, participants additionally received immediate feedback after each problem they solved, which consisted simply of learning whether or not their answer was correct.

One participant was eliminated from analyses for attending law school and an additional six for spending under 10 minutes on the task.

Table 1: Mean scores and perceived scores by condition (standard deviations in parentheses).

	Actual Score	Estimated Score
No Feedback	10.24 (3.77)	8.80 (4.05)
With Feedback	9.57 (4.00)	8.27 (3.96)

Results

Out of the 93 participants included in analyses (55 male, 35 female, 2 other, and 1 unspecified; mean age = 33.33 years), the average completion time was 38 minutes. On average, participants answered 9.90 problems correctly out of 20 (sd = 3.88) and the mean perceived score was 8.54 (sd = 3.99). The difference between actual score and perceived score was deemed significant by a paired-samples t-test ($t(92) = 4.13$, $p < .001$). Table 1 shows that this pattern of underestimation held for both conditions. The overconfidence of the worst performers was limited, presumably given that this test of logical reasoning (from the 2007 LSAT) was significantly more difficult from the logical reasoning test used by of Kruger and Dunning (1999) from the 1993 LSAT. Thus we do not observe the classic Dunning-Kruger effect (overestimation by the worst participants).

Pre- and post- self-assessments were correlated with one

standard deviation of the mean on θ_p, σ_θ.

[4]Link to problems: www.lsac.org/docs/default-source/jd-docs/sampleptjune.pdf

another in the no feedback condition, but not in the feedback condition. In both conditions self-assessments became better correlated with actual performance after completing the assessment (see Table 2), though there was still a pattern of underestimation in the data.

The difference in self-assessment calibration between the conditions was deemed significant by a Fisher r-to-z transformation between the Pearson r values ($z = 4.43$, two-tailed $p < .001$), meaning those in the feedback condition, as anticipated, were much more accurate in estimating their score after the task than those in the no feedback condition.

In a linear model predicting estimated score from true score and condition, a significant regression equation was found ($F(3,89) = 28.62$, $p < .001$ with an R^2 of .48). Specifically, there was no effect of true score, but there were statistically significant effects of condition ($\beta = -4.16$, $p = .014$) and the interaction of true score with condition ($\beta = .41$, $p = .010$), demonstrating that the effect of score on perceived score also depends on the condition, as predicted by our rational model.

To fit the model to the data, we compare model predictions to participants' estimates of their scores relative to their true score. Results from studies of self-assessment have typically organized their data by quartile of performance, as in Figure 3. However, this portrayal of the data eliminates much of its nuance. In the no feedback condition, grouping the self-assessments by true score instead of by quartiles shows more variability (see Figure 4a).

To find the best-fitting parameters for the model given the data, we perform a grid search over μ_θ and ε where we consider values of $\mu_\theta \in [-1,1]$ and $\varepsilon \in [0,0.5]$, taking steps of 0.05. Baseline values were used for the other parameters ($\sigma_\theta = \sigma_\beta = 1$; $\mu_\beta = 0$). The best-fitting model is that with the lowest sum of squared-errors (SSE) between each individual's estimate and the model's prediction. For the no feedback condition, the best fitting model was parametrized by $\varepsilon = 0.4$ and $\mu_\theta = -0.1$ ($SSE = 591.60$), as shown in Figure 4a. Thus participants in the no feedback condition were characterized by low prior perceived ability as well as inaccuracy at estimating their performance on each problem. For the condition with feedback, the best fit model was parametrized by $\varepsilon = 0$ and $\mu_\theta = -0.35$ ($SSE = 145.25$), as seen in Figure 4b. The results from the feedback condition were thus best captured by a model with a low prior on ability and seemingly perfect accuracy guessing performance on each problem, as should be expected given that self-assessments are heavily impacted by people's 'self-concepts' (Ehrlinger & Dunning, 2003).

Table 2: Pearson correlations between pre- and post- self-assessments and between self-assessments and actual score in both conditions.

	Pre/Post	Pre/Score	Post/Score
No Feedback	.67***	.32*	.44**
With Feedback	.03	-.20	.89***

$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

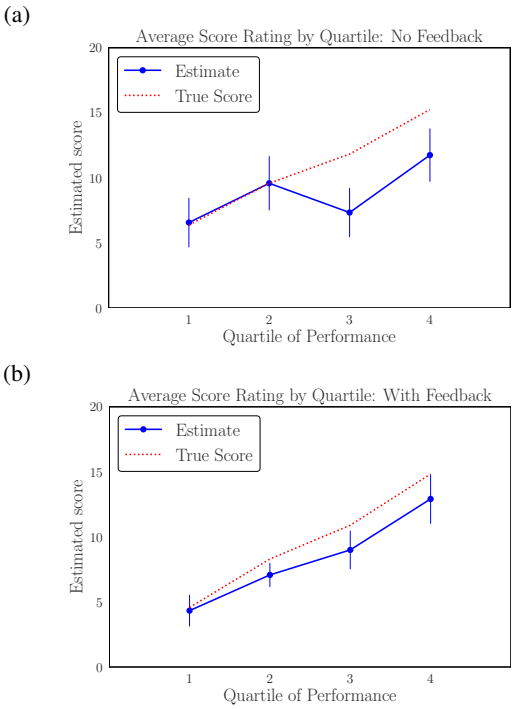


Figure 3: Mean estimates of score (out of 20) by quartile of actual performance in (a) no feedback and (b) with feedback conditions of Experiment 1. Error bars show 95% confidence intervals.

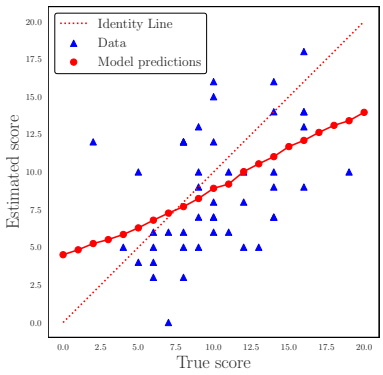
Evaluating Explanations for Dunning-Kruger

The results from the no feedback condition of Experiment 1 were consistent with what was originally found by Kruger and Dunning (1999) and this effect was appropriately attenuated by feedback. However, so far our model has assumed that everyone is equally adept at knowing whether their responses were correct or incorrect, consistent with the regression to the mean hypothesis proposed by Krueger and Mueller (2002). The idea that poor performers are metacognitively impaired in comparison to high performers put forth by Kruger and Dunning (1999) can be captured by extending the model so that there is an ε_p that may differ across individuals in relation to their true ability. Varying ε in relation to true ability expresses the dependence originally asserted by Kruger and Dunning (1999) and allows it to be differentiable from perceived ability θ_p. We make ε_p linearly dependent on person p's score (which serves as a proxy for true ability), with:

$$\varepsilon_p = \varepsilon_0 - \alpha \cdot \frac{\sum_i x_i}{n}, \tag{4}$$

with slope $-\alpha$, intercept ε_0, number of problems n, and $\frac{\sum_i x_i}{n}$ representing the person's scaled score. In the example in Figure 5, we vary ε_p gradually according to Equation 4 with

(a)



(b)

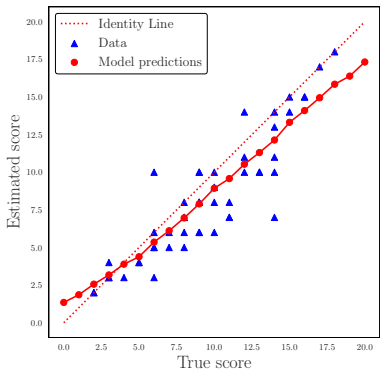


Figure 4: Participants' estimates of score (out of 20) by true score as compared to the best model predictions in (a) no feedback and (b) with feedback conditions of Experiment 1.

$\varepsilon_0 = 0.5$, $\alpha = 0.1$ and then with $\varepsilon_0 = 0.4$, $\alpha = 0.3$ ($n = 10$ in this toy example). This produces greater overestimation at lower true scores, consistent with the pattern Kruger and Dunning (1999) suggested holds for people.

Experiment 2: Comparing Models

We saw in Experiment 1 that our model can capture the typically observed patterns of behavior seen in studies of self-assessment. But to properly evaluate the hypothesis put forth by Kruger and Dunning (1999) that people's self-assessment ability actually *varies* based on their true ability, we need a high-resolution estimate of the form of the function relating true score to estimated score. We replicated the no feedback condition of Experiment 1 with a much larger sample of participants to help determine whether the more complex model is justified by the data.

Methods

Participants A total of 250 participants were recruited on MTurk and were each compensated $3 for their time.

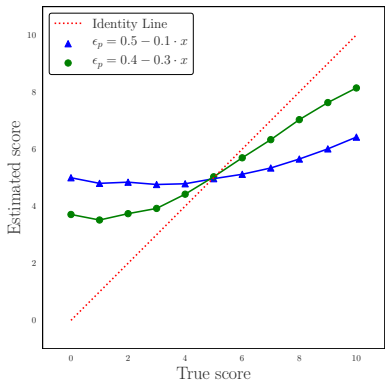


Figure 5: Simulations for the dependent model.

Procedure The procedure was identical to the no feedback condition from Experiment 1. A total of 18 participants were eliminated for spending under 10 minutes and six more for indicating that they had attended law school or taken the LSAT.

Results

Of the 226 participants included in analyses (118 male, 106 female, 1 other, and 1 unspecified; mean age = 36.11 years), the mean completion time was 37 minutes. On average, participants solved 9.39 problems correctly out of 20 (sd = 4.41) and the mean perceived score was 8.10 (sd = 3.89).

Consistent with Experiment 1, pre- and post- self-assessments were correlated with one another ($r = .53$, $p < .001$) and self-assessments became better correlated with actual score after completing the problems (pre: $r = .18$, $p < .01$; post: $r = .56$, $p < .001$). There was also a general pattern of underestimation (see Figure 6).

The version of the original model that minimized the *SSE* was parametrized by $\varepsilon = 0.4$ and $\mu_\theta = -0.15$ ($SSE = 2313.14$), which are very similar to the parameter estimates from the no feedback condition of Experiment 1. Figure 6 depicts the mean self-assessments at each achieved score as compared to this model's predictions for each possible score. There is an increase in overestimation at the low end of the true scores, but this is paralleled by an increase in the variability of the means as relatively few participants performed very poorly. As a consequence the model predictions still fall within the confidence interval for the mean.

We also fit the data to a model with ε dependent on score. The model with the lowest *SSE* had parameters $\mu_\theta = -0.2$, intercept $\varepsilon_0 = 0.45$, and slope $\alpha = 0.1$ ($SSE = 2256.68$). The corresponding plot can be seen in Figure 6. As should be expected as a result of having an additional free parameter, this model gives a closer match to the observed means.

To compare these competing models, we calculated their Bayesian information criteria (*BIC*). The *BIC* for the model with ε dependent on score ($BIC = 1282.26$) was somewhat lower than that of the model with constant ε ($BIC = 1281.60$).

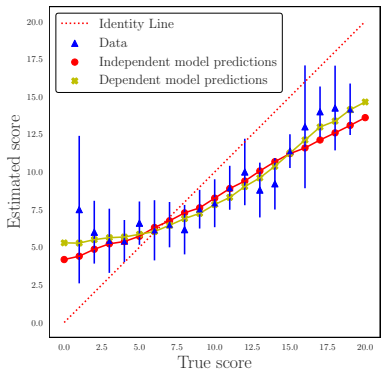


Figure 6: Means of estimated score at each true score in Experiment 2 as compared to best model with constant ε (independent model) where $\varepsilon = 0.4$ and $\mu_\theta = -0.15$ and best dependent model where $\varepsilon_0 = 0.45$, $\alpha = 0.1$, and $\mu_\theta = -0.2$. Error bars show 95% confidence intervals.

Because these models are nested, we also performed a likelihood ratio test yielding $\chi^2(1) = 6.18, p < .05$, which is significant. Though this provides evidence to prefer the more complex model, we acknowledge that there is a very limited amount of data in the tails which is necessary for making a strong conclusion in one direction or the other.

Discussion

We created a formal model to compare competing hypotheses about what causes inaccurate self-assessment of ability. Our model suggests that inaccurate self-assessment is a result of using prior knowledge about one's ability and mistakenly estimating whether one's response was correct or not. We presented two versions of this model: one where participants were equal in their ability to guess whether they solved a problem correctly and one where this skill varied with true ability. Both of these models serve as good approximations of the data (as seen in Figure 6) and comparing these models revealed only weak evidence to prefer the complex model. Future work will explore versions of the model with more possible model parameters (including adjusting the form of the distributions over both θ_p and β_i), which will likely yield more precise parameter estimates. Further studies with larger samples will allow for fully evaluating whether the metacognitive deficit proposed by Kruger and Dunning (1999) can be supported, due to the need for substantial numbers of participants performing at the very low and very high ends.

There are three primary directions for future work. The first is to distinguish between types of self-assessment. For example, the studies presented here also asked participants for their *relative* self-assessments, which is simple enough to apply this model to: rather than converting ability parameter simulations into scores, we can place them relative to one another and observe how these predictions differ from the absolute predictions presented above.

A second avenue to explore is how self-assessments differ across domains. In recent work, Jansen, Rafferty, and Griffiths (2017) found that participants were very accurate in their ability to self-assess after solving algebraic equations. Is this a product of learners generally having more awareness of their ability in math, or is this related to the way in which problems were presented? More awareness of mathematics abilities would be expressed in the model as more accurate priors on ability (μ_θ), while better assessments due to problem presentations would be expressed by low ε, meaning judgments of correctness were more accurate.

Finally, we aim to refine the model by adding complexity and to account for other factors. A lingering question in this literature is whether self-assessment is indeed a cognitive construct. Sitzmann, Ely, Brown, and Bauer (2010) reveal through a meta-analysis that self-assessment measures are more highly correlated with affective outcomes than cognitive outcomes. Of interest in future work will be to explore how affective variables impact results and whether they can be incorporated in a rational model.

We see our results as a first step towards providing a nuanced formalization of how people judge their ability on different types of tasks in a variety of domains. We hope that as our understanding grows we will also develop a better sense of what we do and do not know.

References

Dunning, D., Heath, C., & Suls, J. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(21), 5-17.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Healy, P. J., & Moore, D. A. (2007). Bayesian overconfidence. *http://dx.doi.org/10.2139/ssrn.1001820*.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2017). 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.*

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*(2), 180–188.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing ones own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134.

Kruger, J., & Dunning, D. (2002). Unskilled and unaware–but why? a reply to krueger and mueller. *Journal of Personality and Social Psychology*, *82*(2), 189–192.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, *9*(2), 169–191.