A rational model of sequential 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 (tomg@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544 USA

Abstract

People's assessment of their ability varies in whether it is measured once following a task or sequentially via confidence judgments recorded throughout. Multiple models have been developed to predict one-off judgments of performance, which have often distinguished between peoples' biases about their general ability in a domain and their sensitivity to correctness. We propose a rational model of sequential self-assessment which allows us to make predictions about each individual separately—unlike in the one-off case which looks exclusively at the population level—and to identify, in addition to bias and sensitivity, the extent to which individuals' beliefs are responsive to their most recent evidence over the course of a task. We fit our model to data where participants solve algebraic equations and show that bias, sensitivity, and responsiveness vary meaningfully across participants.

Keywords: Bayesian modeling; Monte Carlo methods; particle filter; self-assessment; metacognition

Introduction

Self-assessment — the act of judging one's performance on a task — is a fundamental metacognitive skill that can be studied in a wide variety of tasks, from trivia (e.g., Burson, Larrick, & Klayman, 2006) to mathematics (e.g., Nelson & Fyfe, 2019). When related to a task where success can be measured, as in an algebra test, self-assessment itself can be measured and analyzed by capturing individuals' beliefs about their performance, and comparing it to their observed performance. Researchers have typically done this in one of two ways, either via one-off judgments made following a task (e.g., in Kruger & Dunning, 1999), or with a sequence of confidence judgments collected throughout a task (used by, e.g., Krueger & Mueller, 2002; Burson et al., 2006). Multiple researchers have sought to model the former type of judgments (e.g., Fleming & Daw, 2017; Healy & Moore, 2007; Krajč & Ortmann, 2008; Jansen, Rafferty, & Griffiths, 2018), which allows for distinguishing group-level differences in bias, or "self-concept" as in Ehrlinger, Johnson, Banner, Dunning, and Kruger (2008), and sensitivity to correctness, both of which are parameters in Fleming and Daw (2017). Previous work has discriminated between these abilities across domains (Jansen, Rafferty, & Griffiths, 2017), task difficulty (Burson et al., 2006), gender (Correll, 2001), and even growth mindset (Ehrlinger, Mitchum, & Dweck, 2016).

Measures of self-assessment that involves sequential confidence judgments made by each individual participant are frequently employed in the educational literature and referred to as "metacognitive monitoring" (e.g., Nelson & Fyfe, 2019). Having people monitor their performance throughout a task gives the chance to identify individual in addition to group-level differences in calibration, and is therefore better for determining what underlies in self-assessment calibration. However, sequential confidence judgments have not been modeled to the same extent as aggregated single judgments.

Here, we present a rational model that makes predictions about individuals' confidence judgments based on (a) their correctness on each problem, (b) their prior beliefs about their ability (frequently referred to as "bias"), and (c) their skill at determining whether they are correct on a single problem (often called "sensitivity"). This approach makes it possible to determine the variability of accuracy in individual-level predictions and also has the potential to reveal individual differences, which can be separately examined and analyzed across different domains. We observe that participants vary greatly in their prior beliefs about their ability. Additionally, some do not seem to know when they have correctly solved a problem while others are quite aware. We additionally see that some individuals are more responsive to their recent progress on a task while others do not update their beliefs as much, so we develop a version of the model that incorporates the idea of responsiveness as a parameter and compare it to models where this parameter is equal to zero.

Using this rational modeling to disentangle causes of miscalibration separately for each person enables us to identify whether there are patterns in responses or personal characteristics (e.g., age or experience learning about a particular domain) that regulate the different parameters in our model. This customized modeling approach promises a more accurate picture of a given individual's metacognitive abilities than any that have been undertaken previously.

Background: Measuring Self-Assessment

Results from studies of self-assessment show that multiple individual-level characteristics may cause differences in calibration to performance on a task. Ehrlinger and Dunning (2003), for example, proposed that a person's "self-concept," their beliefs about their overall skill in a domain, is founda-

tional to their beliefs about their performance on a specific task. These views about the self are likely to be very different, especially in domains like math where variable self-concepts have been widely documented (e.g., Seaton, Parker, Marsh, Craven, & Yeung, 2014). Thus, analyzing individuals' confidence judgments can assist in capturing even smaller individual-level differences across domains.

In Ehrlinger and Dunning (2003), men and women performed comparably on a science test, but women underestimated their ability compared to men. In a similar vein, Correll (2001) argued that cultural beliefs about gender and math ability harmed girls' perceptions of their competence. A model of consecutive self-assessment can still account for these sorts of group-level differences (in additional to individual differences) because this type of analysis will specify the full distribution of individual parameters within groups.

Some have argued that aggregating confidence judgments is a superior method to requesting a single judgment per participant (e.g., Krueger & Mueller, 2002), but really these are different types of judgments that may both be important in distinct ways: single judgments following a task are useful for a person to determine what they will be capable of in the future, while confidence judgments, which convey someone's tracking of their ability, are necessary for determining which more specific skills require targeted study. On a linear equation-solving task, for example, self-assessments made following the task will be used by someone to decide whether to keep practicing at their level or to move on to quadratic equations or another more advanced topic. Tracking performance throughout this task, on the other hand, will provide insight into whether there is a specific algebra skill they are having trouble with (e.g., distribution or combining terms). We will be able to analyze how different individuals update their perceptions of their ability throughout a task with this formal model of sequential self-assessment.

Modeling Sequential Confidence Judgments

Through a computational model, we can generate a more accurate representation of the form of the function that links each participant's confidence throughout a task to their actual performance. In this section, we describe a rational model and the predictions it will make under a variety of circumstances. Following this, we fit alternative versions of the model to data where participants solved algebraic equations.

Model assumptions

Our model makes similar assumptions to previous modeling work (e.g., Jansen et al., 2018; Fleming & Daw, 2017), but at the problem-by-problem level rather than at task completion and treats individuals as making confidence judgments that are consistent with Bayesian inference about their ability. We assume a rational agent makes each judgment based on their beliefs about their ability so far (which includes both their prior beliefs before beginning the task and their performance on already solved problems), the task's difficulty, and individuals' sensitivity to their correctness on each problem.

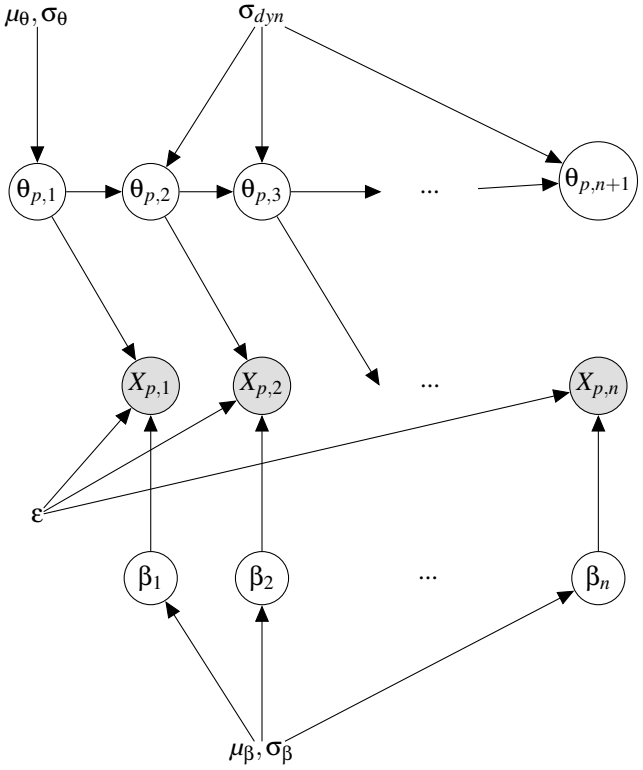

Figure 1: Graphical representation of the model: each observed item $X_{p,t}$ is influenced by latent variables β_t (difficulty of problem at time t) and $\theta_{p,t}$ (perceived ability of person p at time t) as well as a constant ε (ability to determine correctness). The difficulties of all problems β_t and the prior over perceived ability $\theta_{p,1}$ are drawn from normal distributions with means μ_β, μ_θ and standard deviations σ_β, σ_θ. Each subsequent $\theta_{p,t>1}$ is drawn from a $N(\theta_{p,t-1}, \sigma_{dyn})$.

We assume that the priors over a person p's perceived ability before beginning the task ($\theta_{p,1}$) and the difficulty of each problem (β_t) are normally distributed[1] and we compute the probability of a person's correctness at a particular time point t ($X_{p,t}$) which is dependent on perceived ability and difficulty parameters up to and including the current time point.[2]

At each time point, we compute the probability of responding to a problem correctly or not given the person's prior perceived ability and the difficulty of this problem which acts as the *likelihood* in this Bayesian computation. To do so, we use the 1-parameter IRT model, known as the Rasch model. We borrowed this from the psychometrics literature as our representation of the likelihood because it is commonly used to evaluate student ability (Embretson & Reise, 2013). We are turning this idea inward and thinking of people as intuitive psychometricians tracking their own ability. If the problem is

[1] Alternative distributions may be considered in future work, as these would produce different predictions about individuals' beliefs going into the task.

[2] For ease of reading, we drop the p in the subscripts, as we assume that the model is run separately for each individual.

solved correctly, the likelihood is equal to:

$$P(X_t = 1|\theta_t, \beta_t) = \frac{1}{1 + e^{-(\theta_t - \beta_t)}}. \quad (1)$$

When an incorrect response is made, the likelihood is equal to one minus the probability of a correct response: $P(X_t = 0|\theta_t, \beta_t) = 1 - P(X_t = 1|\theta_t, \beta_t) = \frac{1}{1+e^{(\theta_t - \beta_t)}}$.

This version of the likelihood assumes that individuals are flawless in their judgments of correctness, so we include an error parameter (ε), which represents the probability of incorrectly guessing performance on an individual problem:

$$P(X_t = 1|\theta_t, \beta_t, \varepsilon) = (1 - \varepsilon) \cdot P(X_t = 1|\theta_t, \beta_t) + \varepsilon \cdot P(X_t = 0|\theta_t, \beta_t). \quad (2)$$

Because we are modeling sequential confidence judgments, perceived ability at time t depends on perceived ability at all problems up to $t-1$. At time $t = 1$, θ_1 is drawn from a normal distribution with mean μ_θ and standard deviation σ_θ. At all subsequent time points, the dynamics governing how θ_t is related to θ_{t-1} is specified by $p(\theta_t|\theta_{t-1})$. We assume a normal distribution for θ_t centered at θ_{t-1} and that the variance of this distribution, σ_{dyn}, is a parameter of the model which controls how reactive people are to their most recent data.

We additionally need to adjust the likelihood function for all $t > 1$ to incorporate all previous problems, so we define the probability of responding to a question correctly given someone's perceived ability and the difficulty of the problems so far as the product of all likelihoods up through the current problem. We combine this likelihood with a person's previous ability belief $p(\theta_{t-1}, X_{1:t-1})$ and the dynamics of perceived ability $p(\theta_t|\theta_{t-1})$ via Bayes' rule to compute each person's posterior beliefs about their own ability on each problem at time t:[3]

$$p(\theta_t|\theta_{1:t-1}, X_{1:t}) \propto \int_{\beta_k} p(X_k|\theta_k, \beta_k, \varepsilon) p(\beta_k) d\beta_k \cdot p(\theta_{t-1}|X_{1:t-1}) \cdot p(\theta_t|\theta_{t-1}). \quad (3)$$

A graphical representation of the model dependencies is shown in Figure 1. In model simulations presented next, we vary prior beliefs about ability via μ_θ, the likelihood by increasing ε, and the dynamics of perceived ability through changes to σ_{dyn}.

Generating model predictions

Because the integrals in Equation 3 are intractable to calculate exactly, we require an algorithm that can dynamically update the posterior on θ_t in light of new data. We use a standard sequential Monte Carlo method known as a *particle filter* (see Doucet & Johansen, 2009, for an overview). To produce a model simulation for a given set of parameters

[3]To obtain estimates of people's inferences over their ability, we marginalize over the difficulty parameters (β_t).

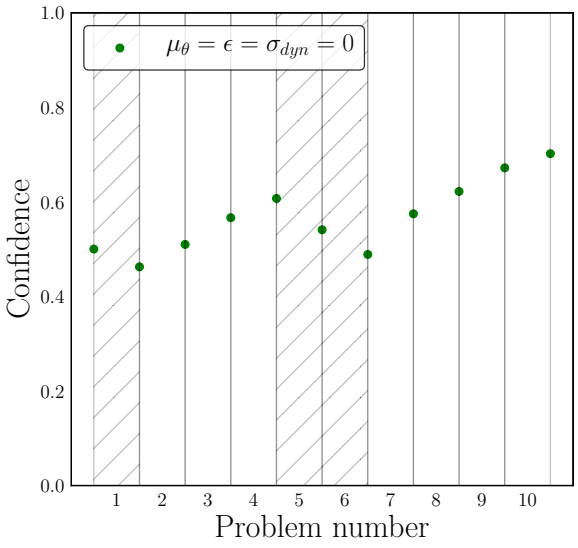

Figure 2: Model predictions in a toy example where participants solve 10 problems in a baseline model (μ_θ, $\mu_\beta = 0$, σ_θ, $\sigma_\beta = 1$, and $\varepsilon = 0$). Each point shows the weighted average of the posterior distribution on each θ_t, corresponding to confidence judgments at each time point. The first judgment made before the first problem is simply based on the prior over ability. The hashed areas demarcate problems solved incorrectly.

with a particle filter, we follow Algorithm 1 to generate posterior distributions of each θ_t given $X_{1:t}$. At each time point $t > 1$, we represent the posterior with a set of n θ_t values, or *particles*, from a probability distribution based on the particles at the previous time $t-1$. Each vector of particles has a normalized set of n weights equal to the cumulative likelihoods as in Equation 3. If the variance of these weights is large we want to remove particles with low weights and multiply ones with higher weights, so we resample n new particles using the normalized weights as a distribution, and then adjust the weights to be uniform. The likelihoods then accu-

Algorithm 1: PARTICLE FILTER ALGORITHM

1. **Sample** a set of n particles θ_t^i, $(i = 1...n)$
   - **(a)** If $t = 0$: from the prior $N(\mu_\theta, \sigma_\theta)$;
   - **(b)** If $t \geqslant 1$: from the particle at time $t-1$ $p(\theta_t^i|\theta_{t-1}^i)$;
2. **Compute weights** $w_t^i(\theta_{1:t}^i)$ which are equal to the product of the previous likelihoods $p(X_{1:t}|\theta_{1:t}^i)$ since resampling: $w_t^i(\theta_{1:t}^i) \propto w_{t-1}^i(\theta_{1:t-1}^i) p(X_t|\theta_t^i)$;
3. **Resample:** if resampling criterion satisfied, resample $\{W_t^i, \theta_{1:t}^i\}$ to obtain n new equally weighted particles $\{\frac{1}{n}, \widehat{\theta_{1:t}^i}\}$

mulate again as new data come in. For each time point, we convert the vector of associated particles into probabilities of a correct response via Equation 1 (which is a sigmoid function) and take the weighted average of all particles to obtain a model prediction of each confidence judgment between 0 and 1 over time (see Figure 2 for a baseline example of model predictions where we set all adjustable parameters ε, μ_θ, and σ_{dyn} equal to zero).[4]

Changing the prior When we adjust the prior over a person's beliefs about their ability (μ_θ), we observe changes to their overall beliefs. In the toy example in Figure 3a, when we assign a higher mean over ability ($\mu_\theta = 1$), confidence judgments tend to be higher overall. Shifting the prior mean downward ($\mu_\theta = -1$) most depresses confidence judgments early on, when the person has limited data from the task, but as they have more experience, their estimates of their ability become more similar to the case with the higher prior mean.

Changing the likelihood Increasing ε to include more error in individual judgments of correctness lowers confidence following correct responses and raises them after incorrect responses (see Figure 3b).

Changing dynamics By varying the dynamics of our model, we can control the extent to which participants learn from their entire set of previous responses. When $\sigma_{dyn} = 0$, a new particle at time t will be exactly the same as the old particle at time $t-1$ because the probability distribution places all the mass at the one location. As σ_{dyn} increases, there is a higher chance of the particle moving farther away from its previous location. In Figure 3c, we observe that a larger value of σ_{dyn} results in recent observations having a greater influence on beliefs. In particular, as seen in this example, when the simulated participant answers multiple problems correctly in a row, their confidence increases more steeply when $\sigma_{dyn} = 2$ than when it is zero (and decreases similarly after multiple sequential incorrect answers).

Fitting the Models to Data

To see how well models with different sets of parameter values compare to actual judgments, we designed an experiment to elicit sequential self-assessments from individuals. To see whether the parameter adjusting the dynamics of the model (σ_{dyn}) is necessary to generate better model predictions, we fit each individual's data to a version of the model with no dynamic updating ($\sigma_{dyn} = 0$)—which we refer to as the *static* model—and a second version, the *dynamic* model, where $\sigma_{dyn} > 0$. This distinction is consistent with the idea of 'mindset' (Ehrlinger et al., 2016) such that the static model captures a fixed mindset (because θ is fixed) while the dynamic model represents a growth mindset (since θ varies and we can track how it changes). We chose a mathematical domain because, as previously discussed, there are many cultural influences on self-views that may impact participants' prior beliefs about their ability (captured by μ_θ), and adult participants will also have had different amounts of math education, which is likely to impact their sensitivity to their correctness (or ε).

Procedure

We conducted a study to elicit confidence judgments from individuals on Amazon's Mechanical Turk. Participants solved 20 multiple-choice algebraic equations and prior to each were asked "You are about to solve a problem. How confident are you that you will solve it correctly?" as well as following the final problem (on a scale from 0 to 100), which resulted in a total of 21 judgments each. All participants received the same problems in the same order, so that problem difficulty was preserved and individuals could be directly compared. To vary problem difficulty, equations required varying amounts of steps and skills (e.g., combining like terms, fractions) to solve, such as $15 - x = 19$ and $6(-10+3x)+2(5x+6/5) = -10x$. There were four multiple-choice options per problem and the three distractor solutions were designed to be the results of different errors a participant might make.

We obtained a total of 199 responses, but excluded 17 for failing an instructional manipulation at the beginning of the survey and an additional 3 who claimed to have used assistive technology to solve the problems (they were asked to use nothing but pencil and paper). This left us with 179 responses for analysis, with an average number of problems correctly solved of 10.85 out of 20.

Model simulations

We compared each individual's data to both the static model (where $\sigma_{dyn} = 0$) and the dynamic model ($\sigma_{dyn} > 0$). In order to make this comparison, we generated sets of model simulations for each participant given their set of correct and incorrect responses by performing a grid search over μ_θ and ε for the static model and these two parameters along with σ_{dyn} for the dynamic model such that values of $\mu_\theta \in [-3,3]$, $\varepsilon \in [0,0.5]$,[5] and $\sigma_{dyn} \in [0.01,6]$ were considered. We took steps of 0.05 for ε, giving 11 possible values, steps of 0.2 for μ_θ, resulting in a total of 31 values, and increasing steps of σ_{dyn}[6] for a total of 31 values, which produced 341 static model predictions and 10,571 dynamic model predictions. These parameter values were chosen based on initial attempts to model a subset of participants such that a representative spectrum of possible parameters were considered.

Results

For each participant, we compared their confidence judgments to each set of model predictions by calculating the sum of squared errors (*SSE*). We took the models with the smallest *SSE* amongst the static models and then the dynamic models to identify the parameters associated with the best fit model

[4] In all models implemented here, we opted to generate $n = 10,000$ particles and used the Effective Sample Size as threshold for determining when to resample.

[5] Because ε is a probability, we only consider values from 0 to 0.5 in our simulations because this value signifies guessing at chance.

[6] The values considered were (0.01, 0.02, 0.03, ..., 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2, 2.5, 3, 4, 5, 6).

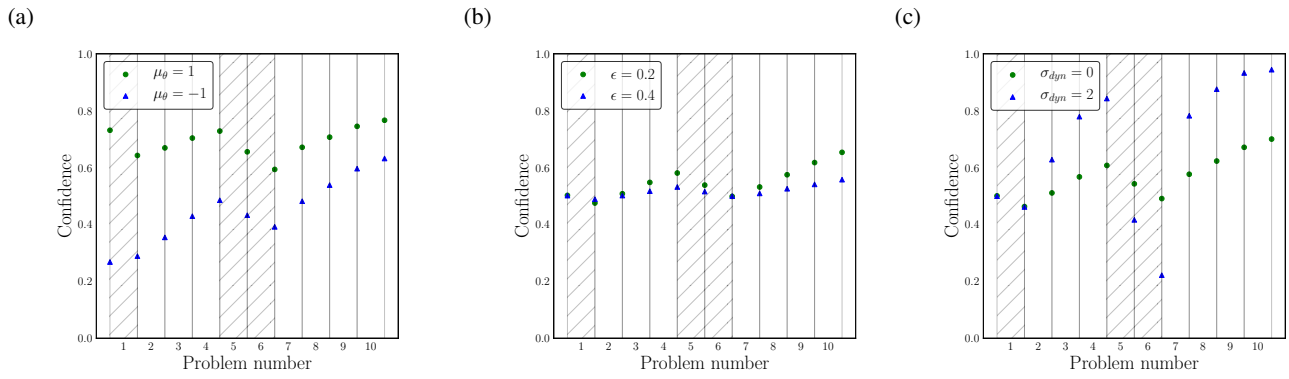

Figure 3: Model predictions on the same toy example as in Figure 2 for (a) when the mean on ability (θ_{pt}) is adjusted ($\mu_\theta = 1$ or -1), (b) when the error parameter ε is adjusted ($\varepsilon = 0.2$ or 0.4), and (c) when (σ_{dyn}) is zero or nonzero ($\sigma_{dyn} = 2$). In all three plots, the parameters we are not adjusting are equal to zero.

in each case. As expected, individuals' confidence judgments were represented by different combinations of parameter values (see Figure 4 for some interesting examples) and though there were some parameters that were more common amongst participants, values varied across individuals (see Figure 5). To determine whether the dynamic or static model fit the data better in general, we calculated each model's Bayesian information criterion (*BIC*) by taking the sum of all individuals' likelihoods (calculated from their *SSE*s) for each model.

We observed a somewhat higher *BIC* for the dynamic model (32232.71) than for the static model (31124.55) when fitting the models to all participants. Since the models are nested, we conducted a likelihood ratio test yielding $\chi^2(179) = 365.36, p < .001$ which is significantly above the threshold for significance, and thus provides evidence to prefer the dynamic model generally across participants (even though for some, the model with $\sigma_{dyn} = 0$ was a better fit, as seen in Figure 5c.

As we might intuit, it appears that some individuals have more dynamic ability beliefs compared others. We thus calculated both models' *BIC* values for each individual separately since we have 21 confidence judgments per participant: the dynamic model fit the data better for 142 participants (meaning it had a lower *SSE* compared to the static model) and for 35 participants, this difference was significant such that the χ^2 test with one degree of freedom was above the threshold of 3.84 for significance (e.g., $\chi^2(1) = 11.20, p < 0.001$ shown in Figure 4a and $\chi^2(1) = 8.91, p < 0.01$ in Figure 4b). This suggests that sequential self-assessment judgments reflect a changing estimate of underlying ability for many people and that at least on an algebraic equation-solving task, most individuals update their confidence according to performance on the most recent problems. We can clearly see that these parameters return interpretable and meaningful results on an individual rather than group level which contributes information beyond what previous metacognitive modeling efforts have provided.

Discussion

We constructed a Bayesian model to predict confidence judgments made sequentially throughout a task and observed that different parameter values described the best-fit models for individual participants. This confirms real-world intuitions that there are individual differences in prior beliefs about ability (μ_θ), knowledge of performance following a given problem (ε), and reactivity to recent performance (σ_{dyn}). Based on the examples presented, we can also conclude that our rational model can accurately predict individuals' confidence judgments, which acts as a good proxy for evolving perceived ability over time. Our approach demonstrates that, by modeling their moment-by-moment confidence judgments, we can identify individuals with dynamic (growth mindset) or static (fixed mindset) beliefs about their own ability in a given domain without relying on self-report of the construct itself. With this dataset, we were able to see that more dynamic versions of the model generally provided superior model fits to human confidence judgments.

There are many avenues for further expanding upon the model presented in this paper. Next, we will look at individual characteristics that might determine groupings of parameter values. We exclusively analyzed data from an algebraic equation-solving task, so in future work we will compare the range of individual parameters across different domains. Specifically, we will compare metacognitive ability on mathematical tasks to trivia tasks where people may have lower sensitivity to their performance following each item (captured by higher ε values) or potentially increased sensitivity since they cannot make simple computational errors (so lower ε). This will likely depend on the type of trivia asked as well as the question formats.

The data we collected contain not only confidence judgments, but overall judgments of ability following a task. In future work, we will compare versions of the model that predict confidence judgments to versions that predict one-off estimates of performance to see whether in aggregate, peo-

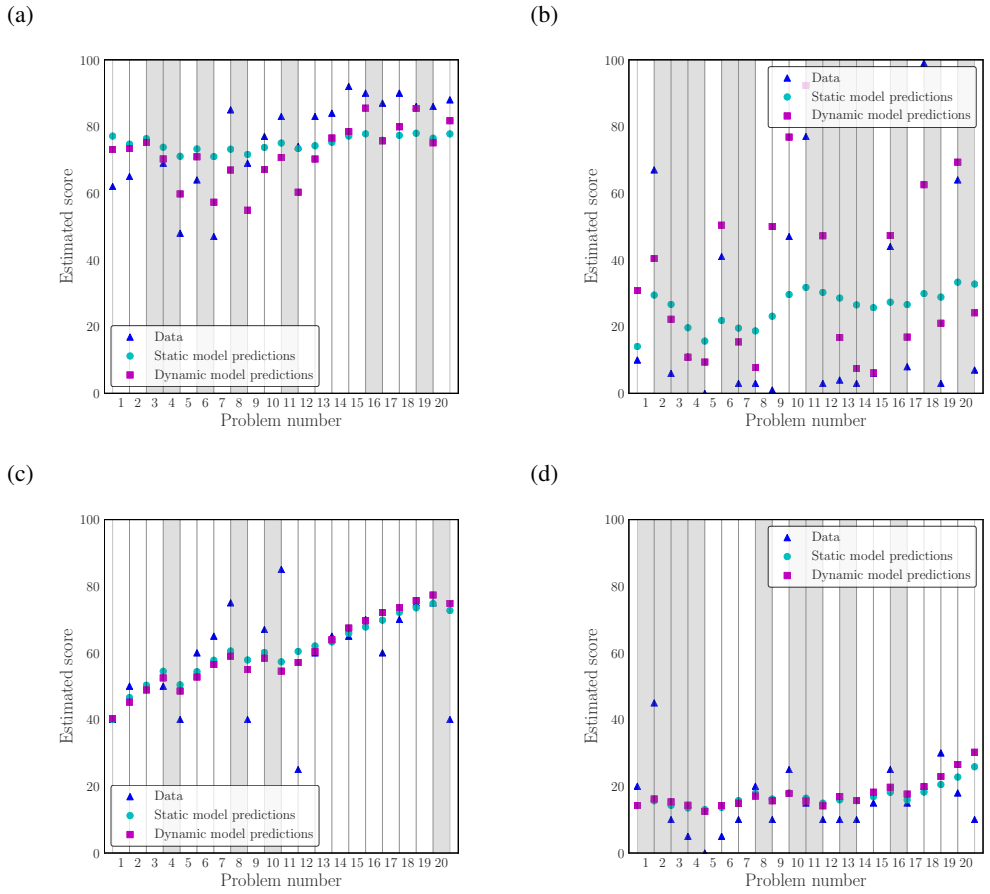

Figure 4: Model predictions for four example participants where grey areas signify incorrect responses: (a) the best fit static model was parametrized by $\varepsilon = 0.25$ and $\mu_\theta = 1$ and the best fit dynamic model had parameters $\varepsilon = 0.2$, $\mu_\theta = 1$, and $\sigma_{dyn} = 0.6$; (b) was fit by $\varepsilon = 0.05$ and $\mu_\theta = -1.7$ while the dynamic model is fit by $\varepsilon = 0.1$, $\mu_\theta = -1.3$, and $\sigma_{dyn} = 2$; (c) the best fit static model for this participant was parametrized by $\varepsilon = 0.3$ and $\mu_\theta = -0.3$ while the dynamic model is fit by $\varepsilon = 0.45$, $\mu_\theta = -0.1$, and $\sigma_{dyn} = 1.6$; (d) was best fit by $\varepsilon = 0.25$ and $\mu_\theta = -1.9$ for the static model and $\varepsilon = 0.35$, $\mu_\theta = -2$, and $\sigma_{dyn} = 0.2$ for the dynamic model. The dynamic model fit the data significantly better than the static model in (a) and (b). In (c) and (d), the dynamic and static models fit the data equally well. Confidence judgments were made out of 100, so displayed are scaled versions of participants' responses here.

ple are better calibrated to their ability during or after a task and be able to inquire about what is behind someone having greater accuracy in their assessments at one time point and not another. Individuals might be very knowledgeable in their confidence judgments, for example, but then overestimate their performance following the task to maintain their self-concept. There might be similarly protective instincts in the reverse direction, where people may be optimistic during a challenging task to maintain their motivation, but end up with a realistic picture of their ability following the task.

Given how frequently metacognitive monitoring is studied in educational domains, we hope to apply this rational model of self-assessment to more real-world data in the hopes of gleaning what is at the source of metacognitive judgments for students on an individual basis and demonstrate the usefulness of computational modeling in more applied settings.

Acknowledgments

This work was made possible thanks to an NSF Graduate Research Fellowship to RJ and a Templeton World Charity Foundation grant to TG for $199,707.

References

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60–77.

Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American journal of Sociology*, *106*(6).

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, *12*(656-704), 3.

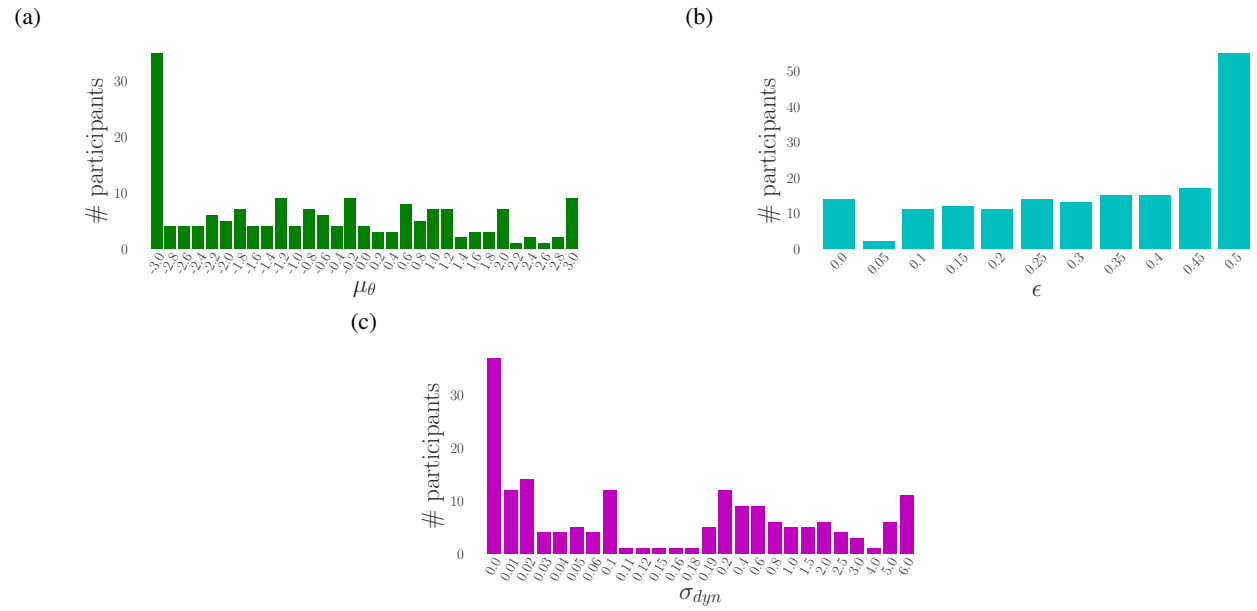

Figure 5: Histograms of values of (a) μ_θ, (b) ϵ, and (c) σ_{dyn} for each person's best-fit model. Note that when $\epsilon = 0$ this indicates that the static model fit the data better for that individual.

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.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121.

Ehrlinger, J., Mitchum, A. L., & Dweck, C. S. (2016). Understanding overconfidence: Theories of intelligence, preferential attention, and distorted self-assessment. *Journal of Experimental Social Psychology*, *63*, 94–100.

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.

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*.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2018). Modeling the Dunning-Kruger effect: A rational account of inaccurate self-assessment. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Krajč, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, *29*(5), 724–738.

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 one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121—1134.

Nelson, L. J., & Fyfe, E. R. (2019). Metacognitive monitoring and help-seeking decisions on mathematical equivalence problems. *Metacognition and Learning*, *14*(2), 167–187.

Seaton, M., Parker, P., Marsh, H. W., Craven, R. G., & Yeung, A. S. (2014). The reciprocal relations between self-concept, motivation and achievement: juxtaposing academic self-concept and achievement goal orientations for mathematics success. *Educational psychology*, *34*(1), 49–72.