What Determines Human Certainty?

Louis Martí (lmarti13@gmail.com)
Francis Mollica (mollicaf@gmail.com)
Steven Piantadosi (spiantadosi@gmail.com)
Celeste Kidd (celestekidd@gmail.com)
Department of Brain and Cognitive Sciences, University of Rochester,
Rochester, NY 14627 USA

Abstract

Previous work on concept learning has focused on how concepts are acquired without addressing metacognitive aspects of this process. An important part of concept learning from a learner's perspective is subjectively knowing when a new concept has been effectively learned. Here, we investigate learners' certainty in a classic Boolean concept-learning task. We collected certainty judgements during the concept-learning task from 552 participants on Amazon Mechanical Turk. We compare different models of certainty in order to determine exactly what learners' subjective certainty judgments encode. Our results suggest that learners' certainty is best explained by local accuracy rather than plausible alternatives such as total entropy or the maximum a posteriori hypothesis of an idealized Bayesian learner. This result suggests that certainty predominately reflects learners' performance and feedback, rather than any metacognition about the inferential task they are solving.

Keywords: Concepts; metacognition; learning; human experimentation; symbolic computational modeling; certainty; ideal learning model

Introduction

Most of us are certain that we landed on the moon, but many of us are far less certain about who will win this year's presidential election. Ideally, our certainty would be a direct reflection of the evidence we observe, but is our sense of certainty actually calibrated to reality? Several studies have demonstrated that individuals presented with disconfirming evidence can become even more entrenched in their original beliefs. Tormala and Petty (2004, 2011) found that when individuals were confronted with messages that they perceived to be strong (e.g., from a expert source) but went against their existing beliefs, their certainty regarding those beliefs *increased* instead of decreased. In contrast, within the visual domain, there is evidence that individuals not only calculate their own subjective measure of visual uncertainty, but that their subjective uncertainty is predictive of objective uncertainty (Barthelme & Mamassian, 2009). In other words, individuals' certainty of visual stimuli reflects veridical probabilities.

The Dunning-Kruger effect provides further evidence of a miscalibration between reality and certainty (Dunning & Kruger, 1999). Dunning and Kruger demonstrated a metacognitive inability of unskilled individuals to recognize their own incompetence. This results in an inflated sense of certainty regarding their own performance and aptitude. The inverse was also found, in which highly competent individuals would be less certain regarding their own abilities in relation to others.

Here, we test whether individuals' subjective certainty while acquiring novel concepts is driven by objective probabilities. Are learners as certain as they should be given the data, or is their subjective sense of certainty driven by other factors (e.g., accuracy, quantity of observed data)? This question has implications for understanding the subjective experiences that accompany concept discovery, which may themselves interact with future learning.

Boolean concept learning as a prototypical domain

Historically, Boolean concept-learning tasks have been used to study concept acquisition because they allowed researchers to study the mechanisms of the learning process in a simplified domain with a known, limited hypothesis space (e.g., Bruner, Goodnow, & Austin, 1967; Feldman, 2000; Goodman et al., 2008; Shepard, Hovland, & Jenkins, 1961).

As an example, consider a classic Boolean concept-learning paradigm. In this task, participants are asked to respond *yes* or *no* to a series of images and are given feedback after each response. Each image has a shape, size, and color that has one of two values, resulting in a total of eight different images. The accuracies of the participants' *yes* or *no* responses are determined by an unstated concept such as "large black square" or "triangle" that subjects must discover. This latent concept is a Boolean rule which can easily be stated in logic, meaning that it is straightforward to quantify representations that learners are likely to be using. Feldman (2000) tested participants on 41 different concepts spread across six families corresponding to the number of positive examples and feature dimensions used in each stimulus. His results showed that performance in learning decreased as Boolean complexity increased, indicating that concept difficulty was directly proportional to the number of Boolean operators in the shortest logically equivalent expression. For example, the concept "large red triangle" should be more difficult than "red triangle" since the former incorporates an additional feature. The idea of simplicity-driven concept learning was put into a probabilistic setting by Goodman et al. (2008), who constructed a Bayesian learner that tried to acquire concepts h from data d according to an idealized model of $P(h \mid d)$. The prior in this model favored simplicity, and the likelihood favored hypotheses that explained the observed labels. In this way, it was able to combine a formalization of a simplicity preference with Bayesian inference from the observed data, providing a close fit to empirical learning.

These tasks provide an ideal domain for us to study certainty since there exist well-tested models and theories of the processes that guide learning in such simple domains. We collect subjective measurements of learners' certainty or confidence throughout a novel concept-learning task. We then compare learners' self-reported certainty throughout the task to a set of models we constructed to represent several a priori plausible, formal theories about what quantitative measure subjective certainty might reflect. We constructed a rule learning model similar to Goodman et al. (2008) in order to provide an idealized quantification of formal measures of certainty, which we then compare to human judgements.

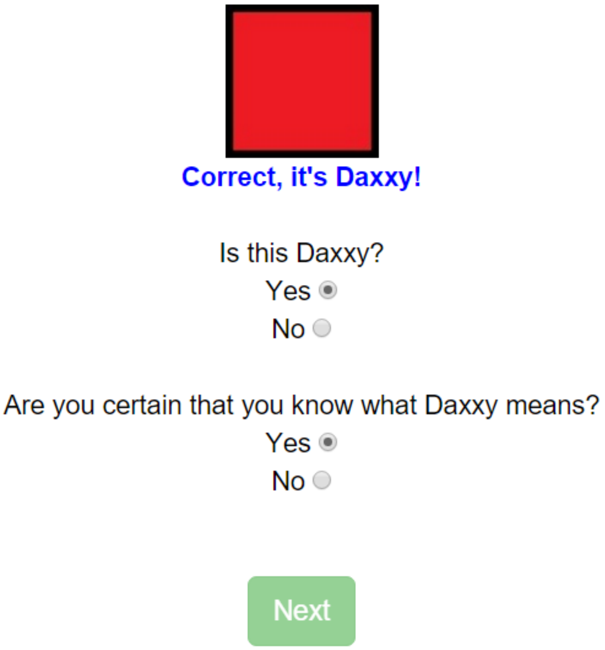
Methods

We tested participants in a novel concept-learning task during which we measured their knowledge of the concept (via *yes* or *no* responses) and their certainty throughout the learning process.

We recruited 552 participants on Amazon Mechanical Turk. Participants clicked to consent to the study before viewing the task instructions. The instructions explained that the participant's task was to figure out the meaning of a word that represented a certain concept. Participants practiced on eight practice trials to ensure that they understood the task before proceeding to the actual study. During the practice trials, participants saw either a cat or a dog, and had to guess whether each item fit the undisclosed concept for a novel word or not by responding *yes* or *no*. In addition to guessing, participants had to report whether or not they were certain about the concept for the novel word. After each guess, participants received feedback about whether or not their guess was correct. For the practice trials, the novel word always referred to the concept of "cat".

For the experimental trials (see Figure 1), participants saw one of ten conditions, each composed of of 24 trials. Each condition represented one unique concept, such that each participant made judgements for only one concept. In a partial replication of the Shepard et al. study (1961), the observations spanned three binary dimensions: shape (square or triangle), color (red or green), and size (large or small). A total of eight images were used across all conditions which exhaustively spanned the space. Across all conditions participants would see these eight images in blocks of three with the ordering of the images assigned randomly per condition. Each condition tested for a different concept with varying complexity (see Table 1).

Concepts 1, 5, 6, 7, 8 and 9 (Table 1) were identical to concepts used in both the Shepard et al. (1961) and Feldman (2000) experiments. These concepts exhaustively spanned across the concept family consisting of three features and four positive examples. Additional conditions were added to test for potential differences between operators. Concept 9—"*(green and large and triangle) or (green and small and square) or (red and large and square) or (red and small and triangle)*"—is predicted to be the most difficult as it essentially transforms that condition into a rote memorization task.



Correct, it's Daxxy!

Is this Daxxy?

Yes ●

No ○

Are you certain that you know what Daxxy means?

Yes ●

No ○

Figure 1: Participants saw 24 trials (as above) in sequential order, randomized between-conditions. After responding, feedback was displayed for one second (if correct) or two seconds (if incorrect) before the next stimuli was shown. Previous feedback was not displayed at any time.

Like the practice trials, each participant gave a "yes" or "no" answer as to whether the current image was part of the concept. On each trial, each participant also indicated whether they were certain regarding their answer, providing a binary forced choice judgment. After their responses, participants received feedback on their responses. Correct responses received one second of feedback before the next trial commenced. Incorrect responses were penalized with a slower two seconds of feedback before the next trial to incentivize attention to the task.

Ideal learning model

We aim to address the question of whether learners' subjective sense of certainty reflects veridical probabilities. In other words, do learners feel as certain as is justified by the observed data? Addressing this question requires us to use an ideal learning model in order to determine how confident an ideal learner should be (given the uncertainty of the model). Here we use an ideal learning model that has already been used to formalize concept learning in a probabilistic setting in which notions of certainty and uncertainty (e.g., Shannon, 1948) are well defined. Though ideal learning models of this type have been used before to understand novel concept acquisition, they have not previously been applied towards understanding learners' subjective sense of certainty.

The ideal learning model was developed using Python

		Concept
1	RED	red
2	AND	red and small
3	OR	red or small
4	XOR	red xor small
5	AND OR AND	(red and small) or (green and large)
6	Complex 1	(green and large and triangle) or (green and large and square) or (green and small and triangle) or (red and large and square)
7	Complex 2	(green and large and triangle) or (green and large and square) or (green and small and triangle) or (red and large and triangle)
8	Complex 3	(green and large and triangle) or (green and large and square) or (green and small and triangle) or (red and small and square)
9	Memorization	(green and large and triangle) or (green and small and square) or (red and large and square) or (red and small and triangle)
10	XOR XOR	red xor small xor square

Table 1: Concepts presented to subjects in the experiment.

Rule
START → PREDICATE
START → TRUE
START → FALSE
PREDICATE → and(PREDICATE, PREDICATE)
PREDICATE → or(PREDICATE, PREDICATE)
PREDICATE → not(PREDICATE)
PREDICATE → red(x)
PREDICATE → green(x)
PREDICATE → triangle(x)
PREDICATE → square(x)
PREDICATE → large(x)
PREDICATE → small(x)

Table 2: Grammar used to generate logical rules in the idealized learning model. The variable x is the current object.

and the Language Of Thought library, *LOTlib* (Piantadosi, 2014). This model defines a probabilistic context-free grammar (PCFG) with a set of primitives: red, green, triangle, square, large, and small, and logical operations (shown in Table 2). The PCFG serves as a prior over hypotheses and specifies an infinite hypothesis space.

To establish a tractable hypothesis space, the model drew 1,000,000 samples from the posterior distribution of hypotheses (i.e., hypotheses scored by simplicity and fit to the data) using tree-regeneration Metropolis-Hastings (Goodman et al., 2008) and stored the best 1,000 hypotheses at each data amount that subjects saw. The model incorporated parameters for the noise in the data (alpha) and a power law memory decay on the likelihood of previous data[1] (beta), best fit as 0.64 and 0 respectively.

Analysis

We considered and compared several different models of what might drive uncertainty. Perhaps the most natural is that learners might use the uncertainty of an idealized learning model, as quantified by the posterior **entropy** (Shannon 1948). This provides a measure of the number of bits of information learners have yet to discover about which hypothesis is the true generator of the data. However, it also may be the case that learners tend to pick probable hypotheses, and their uncertainty reflects only the probability of the best hypothesis. We refer to this as the **MAP** model. We may also consider **maximum likelihood** in which the prior is ignored, corresponding to a maximum likelihood model over the structured, compositional hypothesis space. Beyond these idealized model-based theories, it is critical to include a variety of trial-level alternatives. It could be for instance that participants just become more confident as they complete more of the experiment, a model we refer to as **Trial**. Alternatively, certainty may just reflect a measure of their performance so far, reflecting a lack of objective self-awareness of how much certainty they should have. **Total Accuracy** quantifies performance on all previous trials of the experiment. We also include **Local Accuracy** measures of how well subjects have done on the previous N trials ($N = 2,3,4,5$), potentially incorporating their performance on the current trial (e.g. the one they are responding to), called **Local Accuracy Current**. If this predictor beat out the others, it would indicate that learners are only certain when they anticipate being able to guess accurately on the current trial. The **Current Accuracy** model is a baseline that simply quantifies whether participants were right on the next trial. Its performance as a predictor shows whether subjective certainty is well-calibrated to true accuracy on the next item, regardless of the underlying computational processes.

Logarithmic transformations are common in psychophysics (Stevens, 1957). Therefore, each of these predictors was considered in its standard form, as well as under a logarithmic transformation, yielding a total of 32 models. The accuracy predictors used a $log(1+x)$ transformation to avoid problems with zeroes.

Results

Certainty and accuracy by concept

We composed and evaluated plots of participants' certainty and accuracy over the course of the experiment for each concept in order to determine *(1)* whether certainty and accuracy improved over the course of the experiment, *(2)* whether the-

[1] Weighting the log likelihood of an example n back by $(n+1)^{-\beta}$.

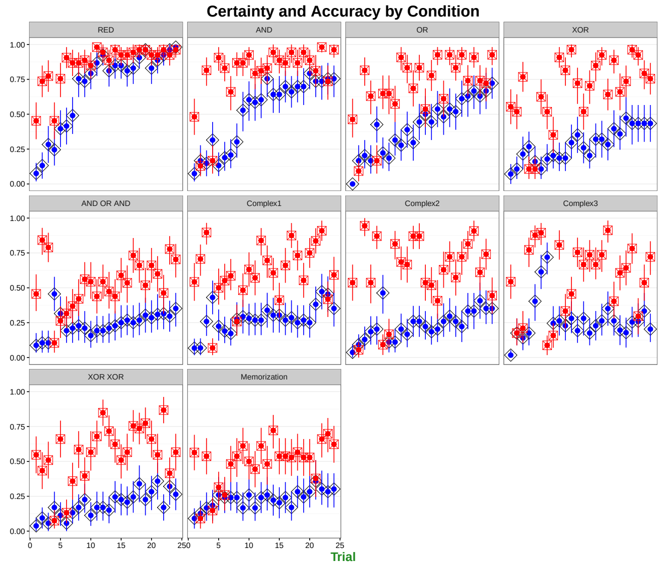


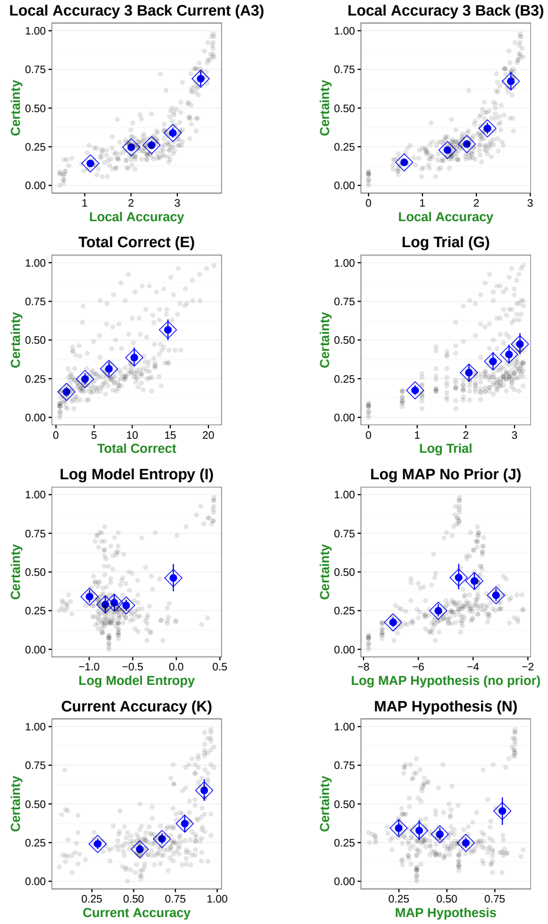
Figure 2: Mean certainty (blue) and mean accuracy (red) across concept conditions

oretically harder concepts (according to Feldman 2000) were, in fact, more difficult for participants, and *(3)* whether participants' certainty correlated with their accuracy in general.

Figure 2 shows participants' certainty and accuracy (*y*-axis) over trials of the experiment (*x*-axis). The increasing trend of the accuracy curves reaches ceiling for some concepts, indicating that participants successfully acquired them. In other conditions, participants did not reach ceiling, indicating that they did not acquire the target concept. This is actually beneficial to our analysis as it allows us to analyze conditions and trials in which participants should have high uncertainty. The certainty curves follow a generally increasing trajectory, but only reach high values (ceiling probability of a participant reporting being certain) in conditions in which participants also achieved high accuracy. The increasing trend of certainty in conditions for which accuracy does not go above 50% may be reflective of overconfidence.

Predictors of certainty

Figure 3 shows certainty (*y*-axis) over several different key predictors of certainty (*x*-axis). Local accuracy models have low dispersion, meaning that individuals with low local accuracy have low certainty and individuals with high local accuracy are highly certain. There are no cases where an individual is highly accurate and highly uncertain and no cases where an individual has low accuracy and is highly certain. On the other hand, total correct and log trial are highly linear but have high dispersion. This is likely due to condition effects. In conditions where the concept is extremely simple (e.g. ”red”) participants might reach high certainty extremely quickly and, due to a low level of negative feedback (incorrect responses), remain highly confident. Current accuracy has a similar shape and dispersion but is not a good predictor due to situations in which the participant is very uncertain but happens to get the trial correct by chance. Finally, log model entropy and the MAP plots perform very poorly. Data points are completely scattered and the predictors perform poorly because of this. For example, although there are many likely MAP hypotheses for which participant certainty is high, there are just as many for which certainty is low. This could be due to complicated concepts for which the ideal learner model does well but participants do not.



(a) (b)

Figure 3: Visualizations of several key model fits, giving the participant response means for each concept and trial (gray) and binned model means in each of five quantiles (blue) for certainty rating (*y*-axis) as a function of model (*x*-axis). Straight lines with low variance correspond to models which accurately capture human performance.

Model comparison results

Before examining how well the certainty models predict human certainty, it is important to verify that the ideal learning model predicts human accuracy during the task. A logistic regression predicting behavioral accuracy from model accuracy

	Model	AIC	Pseudo R^2	Log Likelihood	Beta	Standard Error
A3	Local Accuracy 3 Back Current	14753.0	0.11	-7374.5	1.11	0.03
A4	Local Accuracy 4 Back Current	14778.8	0.11	-7387.4	0.85	0.02
B3	Local Accuracy 3 Back	14786.6	0.11	-7391.3	1.33	0.04
B4	Local Accuracy 4 Back	14803.0	0.11	-7399.5	0.99	0.03
A5	Local Accuracy 5 Back Current	14816.3	0.11	-7406.1	0.68	0.02
B5	Local Accuracy 5 Back	14838.8	0.11	-7417.4	0.76	0.02
A2	Local Accuracy 2 Back Current	14872.6	0.11	-7434.3	1.44	0.04
B2	Local Accuracy 2 Back	14892.2	0.11	-7444.1	1.91	0.05
C3	Log Local Accuracy 3 Back Current	14948.6	0.10	-7472.3	3.40	0.10
D3	Log Local Accuracy 3 Back	14990.2	0.10	-7493.1	3.24	0.10
C4	Log Local Accuracy 4 Back Current	15010.9	0.10	-7503.5	2.92	0.09
C2	Log Local Accuracy 2 Back Current	15018.9	0.10	-7507.4	3.78	0.11
D2	Log Local Accuracy 2 Back	15037.6	0.10	-7516.8	3.83	0.11
D4	Log Local Accuracy 4 Back	15051.6	0.10	-7523.8	2.73	0.08
C5	Log Local Accuracy 5 Back Current	15078.1	0.09	-7537.0	2.53	0.08
D5	Log Local Accuracy 5 Back	15121.0	0.09	-7558.5	2.32	0.07
A1	Local Accuracy 1 Back Current	15215.5	0.09	-7605.8	1.88	0.05
C1	Log Local Accuracy 1 Back Current	15338.3	0.08	-7667.1	3.94	0.12
B1	Local Accuracy 1 Back	15346.5	0.08	-7671.3	2.92	0.09
E	Total Correct	15389.5	0.08	-7692.7	0.14	0.00
D1	Log Local Accuracy 1 Back	15430.1	0.07	-7713.1	4.32	0.14
F	Log Total Correct	15525.0	0.07	-7760.5	0.77	0.03
G	Log Trial	15911.6	0.04	-7953.8	0.70	0.03
H	Trial	16014.1	0.04	-8005.1	0.07	0.00
I	Log Entropy	16205.5	0.03	-8100.8	-1.05	0.05
J	Log Maximum likelihood	16207.4	0.03	-8101.7	0.30	0.02
K	Current Accuracy	16339.6	0.02	-8167.8	0.69	0.04
L	Log Current Accuracy	16339.6	0.02	-8167.8	1.00	0.06
M	Entropy	16389.6	0.02	-8192.8	-0.49	0.03
N	MAP	16545.0	0.01	-8270.5	0.93	0.10
O	Log MAP	16593.9	0.00	-8295.0	0.29	0.04
P	Maximum likelihood	16609.7	0.00	-8302.8	5.02	0.90

Table 3: Performance of predictors in determining subjective certainty. All were significant at $p < .001$.

results in a significant relationship, B = 3.206, $p < .001$.

Table 3 shows the full model results, giving the performance of each model in predicting certainty ratings. These have been sorted by the primary measure of performance, AIC, which quantifies the fit of each model penalizing its number of free parameters (closer to $-\infty$ is better). In this case, the AIC is simply the AIC score of a logistic regression including the variable of interest. This table also provides a pseudo R^2 measure, giving a rough measure of the "amount of variance" accounted for by each model (this is not literally an R^2 since amount of variance does not have a clear analog in logistic models). The reported numbers also include a β giving the regression coefficient, its standard error, and a two-tailed p-value comparing it to zero.

As this table makes clear, the **Local accuracy** models outperform any of the alternatives, a pattern which is robust to the way in which local accuracy is quantified (e.g. the number back that are counted or whether the current trial is included). The quantitatively best model A3 tracks accuracy over the past three trials and includes the future accuracy on the next trial. One possible explanation for this is that individuals are simply deciding their own certainty based on recent performance and whether or not they think they know the answer to the current item.

Interestingly, model E, the **Total Correct** count of responses, is the second best predictor of certainty outside of the local accuracy models. The high performance of this and local accuracy models implies that people's certainty is largely influenced by their own perception of how well they are doing on the task. It is important to note that all the models which use either local accuracy or total correct outperform all other models.

The relative poor performance of the number of **Trial**s in predicting performance (model G) indicates that participants are not simply becoming more certain over time regardless of performance. It also excludes the possibility that learners are waiting for exhaustive data before becoming certain. If they were, a sudden spike of certainty would be visible at trial 8, when in our experimental design they have seen all possible feature dimensions and outcomes.

Strikingly, the poor performance of of the **Entropy** and **MAP** models rules out that subjective certainty is calibrated with an ideal learner. The poor performance of these models is consistent with the theory that learners are likely not maintaining more than one hypothesis in mind—perhaps they store a sample from the posterior, but do not have access to the full posterior distribution. Such a failure of metacognition is consistent with the poor performance of **Current accuracy**, a measure of whether or not the participant got the next trial correct. Subjective certainty does not accurately predict accuracy on the current example, or vice versa.

While entropy is not a major predictor of certainty, a generalized linear mixed model fit by maximum likelihood provides evidence that it is a significant factor when controlling for the most predictive model (**Local Accuracy 3 Back Current**), B = 0.097, p = .005. However, despite being significant, the effect of entropy is small, especially when compared to the effect of local accuracy, B = 1.264, p <.001. This evidence does not rule out the possibility of unknown factors fully mediating the relationship between entropy and certainty.

Conclusions and Discussion

Our analyses revealed that local accuracy is the best predictor of certainty in our simple concept-learning task. Further, this effect is robust to exactly how accuracy is computed. This means that participants seem to be basing their certainty on their immediate performance—*inferring certainty from their own behavior and feedback.* Specifically, participants seem to be assessing their performance on the past several items along with a guess on whether or not they know the current item. This general pattern is consistent with metacognitive studies showing that often subjects do not understand—or perhaps even remember—the causes of their own behavior (Johansson et al., 2005; Nisbett & Wilson, 1977). Subjects don't directly observe their own cognitive processes and are often blind to their internal dynamics. This appears to be true in the case of subjective certainty reports. They do not appear to reflect an awareness of how much certainty subjects *should* have.

The analyses also help inform us about which factors *do not* drive certainty, and several of these results are surprising. For example, one reasonable theory of certainty posits that participants could be basing their certainty off of their confidence in the MAP hypothesis under consideration (as in hypothesis-testing accounts of learning). Our analyses do not support this account. If participants were basing their certainty off of the MAP hypothesis that they were considering, the MAP predictors would perform much better. Since the MAP predictors do not form well, it is unlikely that learners' certainty relies on internal estimates of the probabilities that most Bayesian learning accounts assume.

Our analyses also reveal that there is still a lot that we do not understand about human certainty. We tested many major, reasonable hypotheses about the factors that drive human certainty, yet the proportion of variance explained by the highest performing predictor here is only 11%. Thus, although local accuracy performs better than other predictors, it cannot be the whole story. The low performance of the proposed predictors here is surprising given that our hypotheses spanned major hypotheses involving metacognitive awareness of both uncertainty and task performance—factors that most people have previously assumed are major drivers of certainty in learners. Further, these results hint that non-metacognitive factors may play a surprisingly substantial role in influencing human certainty. As an example, neurochemical changes in the brain induced by stimulants such as Adderall and Ritalin are known to robustly influence self-reported confidence on performance in cognitive tasks without actually boosting objective measures of task performance (Smith & Farah, 2011). It is possible that a large component of certainty could reflect factors that are almost entirely removed from the veridical probabilities, such as the context of the judgement or differences in individual learners' overall self-confidence or mood.

Acknowledgments

The authors would like to thank the Kidd Lab, and the Computation and Language Lab for providing feedback.

References

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Comput Biol*, *5*(9), e1000504–e1000504.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Piantadosi, S. T. (2014). *LOTlib: Learning and Inference in the Language of Thought.* available from https://github.com/piantado/LOTlib.

Shannon, C. (1948). A mathematical theory of communities. bell.. 8)/st. *Techn. J*, *27*, 379–423.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Smith, M. E., & Farah, M. J. (2011). Are prescription stimulants smart pills? the epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological bulletin*, *137*(5), 717.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Tormala, Z. L., Clarkson, J. J., & Henderson, M. D. (2011). Does fast or slow evaluation foster greater certainty? *Personality and Social Psychology Bulletin*, *37*(3), 422–434.

Tormala, Z. L., & Petty, R. E. (2004). Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, *14*(4), 427–442.