Journal: **OPEN MIND**

Certainty is Primarily Determined by Past Performance during Concept Learning

Louis Martí[1], Francis Mollica[1], Steven Piantadosi[1], and Celeste Kidd[1]

[1] Brain and Cognitive Sciences, University of Rochester, Rochester, USA

Keywords: certainty, confidence, metacognition, learning, concepts

Abstract

Prior research has yielded mixed findings on whether learners' certainty reflects veridical probabilities from observed evidence. We compared predictions from an idealized model of learning to humans' subjective reports of certainty during a Boolean concept learning task in order to examine subjective certainty over the course of abstract, logical concept learning. Our analysis evaluated theoretically motivated potential predictors of certainty to determine how well each predicted participants' subjective reports of certainty. Regression analyses that controlled for individual differences demonstrated that despite learning curves tracking the ideal learning models, reported certainty was best explained by performance rather than measures derived from a learning model. In particular, participants' confidence was driven primarily by how well they observed themselves doing, not by idealized statistical inferences made from the data they observed.

INTRODUCTION

Daily life requires making judgments about the world based on inconclusive evidence. These judgments are intrinsically coupled to people's subjective *certainty*, a metacognitive assessment of how accurate judgments are. While it is clear certainty impacts behavior, we do not fully understand how subjective certainty is linked to objective, veridical measures of certainty or probability. For example, people presented with disconfirming evidence can become even more entrenched in their original beliefs. Tormala, Clarkson, and Henderson (2011); Tormala and Petty (2004) found that when people were

Corresponding author: Louis Martí, lmarti13@gmail.com

confronted with messages that they perceived to be strong (e.g., from an expert) but contradicted their existing beliefs, their belief certainty *increased* instead of decreased. The Dunning-Kruger effect—by which unskilled people overestimate their abilities and highly competent people underestimate them—also provides evidence of a miscalibration (Kruger & Dunning, 1999). Confidence is also influenced by social factors. Specifically, individuals calibrate their confidence to the opinions of others, irrespective of the accuracy of those opinions (Yaniv, Choshen-Hillel, & Milyavsky, 2009). Tsai, Klayman, and Hastie (2008) found that presenting individuals with more information raised their confidence irrespective of whether accuracy increased. Miscalibration is also present during "wisdom of the crowds" tasks. When questions require specialized information, individuals were equally as confident regardless of accuracy. This applies to both answers to questions and predictions about the accuracy of others (Prelec, Seung, & McCoy, 2017). Additionally, confidence in a memory has no relationship to whether or not the memory actually occurred (Loftus, Donders, Hoffman, & Schooler, 1989; McDermott & Roediger, 1998). Finally, simply taking prescription stimulants (e.g., Adderall, Ritalin) increases individuals' senses of certainty (Smith & Farah, 2011).

Studies examining perceptual phenomena, however, imply a tight link between certainty and reality. Individuals calculate their own subjective measure of visual uncertainty which has been found to predict objective uncertainty (Barthelmé & Mamassian, 2009). Others have found correlates for subjective certainty such as reaction time, stimuli difficulty, and other properties of the data (Drugowitsch, Moreno-Bote, & Pouget, 2014; Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani, Corthell, & Shadlen, 2014). More evidence demonstrating the linkage between perceptual certainty and reality was presented when Sanders, Hangya, and Kepecs (2016) described a computational model that predicted certainty in auditory and numerical discrimination tasks.

Thus, while our certainty might be a useful guide with regard to perceptual decisions, such as trying to locate a friend yelling for help in the middle of the woods, it may be misleading in higher-level domains, such as deciding whether to see a chiropractor versus a medical doctor. However, no experiment has evaluated quantitatively measured changes in certainty during learning in tasks outside of perception. In ordinary life, evidence accumulation is likely to be less like perceptual learning and more like tasks for which learners must acquire abstract information about more complex latent variables—like rules, theories, or structures. Here, we examine certainty during learning using an abstract learning task with an

infinite hypothesis space of logical rules. We present three experiments that used a Boolean concept-learning task to measure how certain learners should have been, given the strength of the observed evidence. With a potentially overwhelming hypothesis space, is a person's subjective certainty driven by veridical probabilities, or by something else?

Historically, Boolean concept-learning tasks have been used to study concept acquisition because they allowed researchers to examine the mechanisms of learning abstract rules or principles while focusing on a manageable, simplified space of hypotheses (Bruner & Austin, 1986; Feldman, 2000; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Shepard, Hovland, & Jenkins, 1961). Experiment 1 compared measures from an idealized learning model to measures derived from participants' behavior to determine which best matched participants' ratings of certainty. Results suggest that the most important predictor of certainty is people's recent feedback/accuracy, not measures of, for example, entropy derived from the model. Furthermore, a logistic regression with the best predictors demonstrates that most of them provide unique contributions to certainty, implicating many factors in subjective judgments. Experiment 2 tested these predictors when participants were not given feedback. These results show that when feedback is removed, model predictors perform no better than in Experiment 1. Experiment 3 examined participants' certainty about individual trials rather than the overall concept. Similar to Experiment 1, in Experiment 3 people primarily relied on recently observed feedback. Our results show that participants used their overall and recent accuracy—not measured or derived from rule-learning models—to construct their own certainty.

EXPERIMENT 1

Motivation

The aim of Experiment 1 was to measure subjective certainty of participants during concept learning and attempt to predict it using plausible model-based and behavioral predictors. In this experiment, certainty judgments were about what underlying concept (rule) generated the data they saw, as opposed to their certainty about the correct answer for any given trial (see Experiment 3).

Methods

Table 1. Concepts presented to participants. Concepts 1 and 5-9 are the Shepard, Hovland and Jenkins family consisting of three features and four positive examples.

		Concept
1	SHJ-I$_{3[4]}$	red
2	AND	red $\wedge$ small
3	OR	red $\vee$ small
4	XOR	red $\oplus$ small
5	SHJ-II$_{3[4]}$	(red $\wedge$ small) $\vee$ (green $\wedge$ large)
6	SHJ-III$_{3[4]}$	(green $\wedge$ large $\wedge$ triangle) $\vee$ (green $\wedge$ large $\wedge$ square) $\vee$ (green $\wedge$ small $\wedge$ triangle) $\vee$ (red $\wedge$ large $\wedge$ square)
7	SHJ-IV$_{3[4]}$	(green $\wedge$ large $\wedge$ triangle) $\vee$ (green $\wedge$ large $\wedge$ square) $\vee$ (green $\wedge$ small $\wedge$ triangle) $\vee$ (red $\wedge$ large $\wedge$ triangle)
8	SHJ-V$_{3[4]}$	(green $\wedge$ large $\wedge$ triangle) $\vee$ (green $\wedge$ large $\wedge$ square) $\vee$ (green $\wedge$ small $\wedge$ triangle) $\vee$ (red $\wedge$ small $\wedge$ square)
9	SHJ-VI$_{3[4]}$	(green $\wedge$ large $\wedge$ triangle) $\vee$ (green $\wedge$ small $\wedge$ square) $\vee$ (red $\wedge$ large $\wedge$ square) $\vee$ (red $\wedge$ small $\wedge$ triangle)
10	XOR XOR	red $\oplus$ small $\oplus$ square

We tested 552 participants recruited via Amazon Mechanical Turk in a standard Boolean concept-learning task during which we measured their knowledge of a hidden concept (via "*yes*" or "*no*" responses) and their certainty throughout the learning process (see Figure 1 and Table 1). In this experiment, participants were shown positive and negative examples of a target concept "daxxy," where membership was determined by a latent rule on a small set of feature dimensions (e.g. color, shape, size), following experimental work by Shepard et al. (1961) and Feldman (2000). The latent rules participants were required to learn varied across a variety of logical forms. After responding to each item, participants were provided feedback and then rated their certainty on what the word "daxxy" meant. For our analyses we considered and compared several different models of what might drive uncertainty (see Table 2). These predictors can be classified into two broad categories. Model-based predictors were calculated using our ideal learning model while behavioral predictors were calculated using the behavioral data (see Supplemental Materials Appendix A for additional method details).

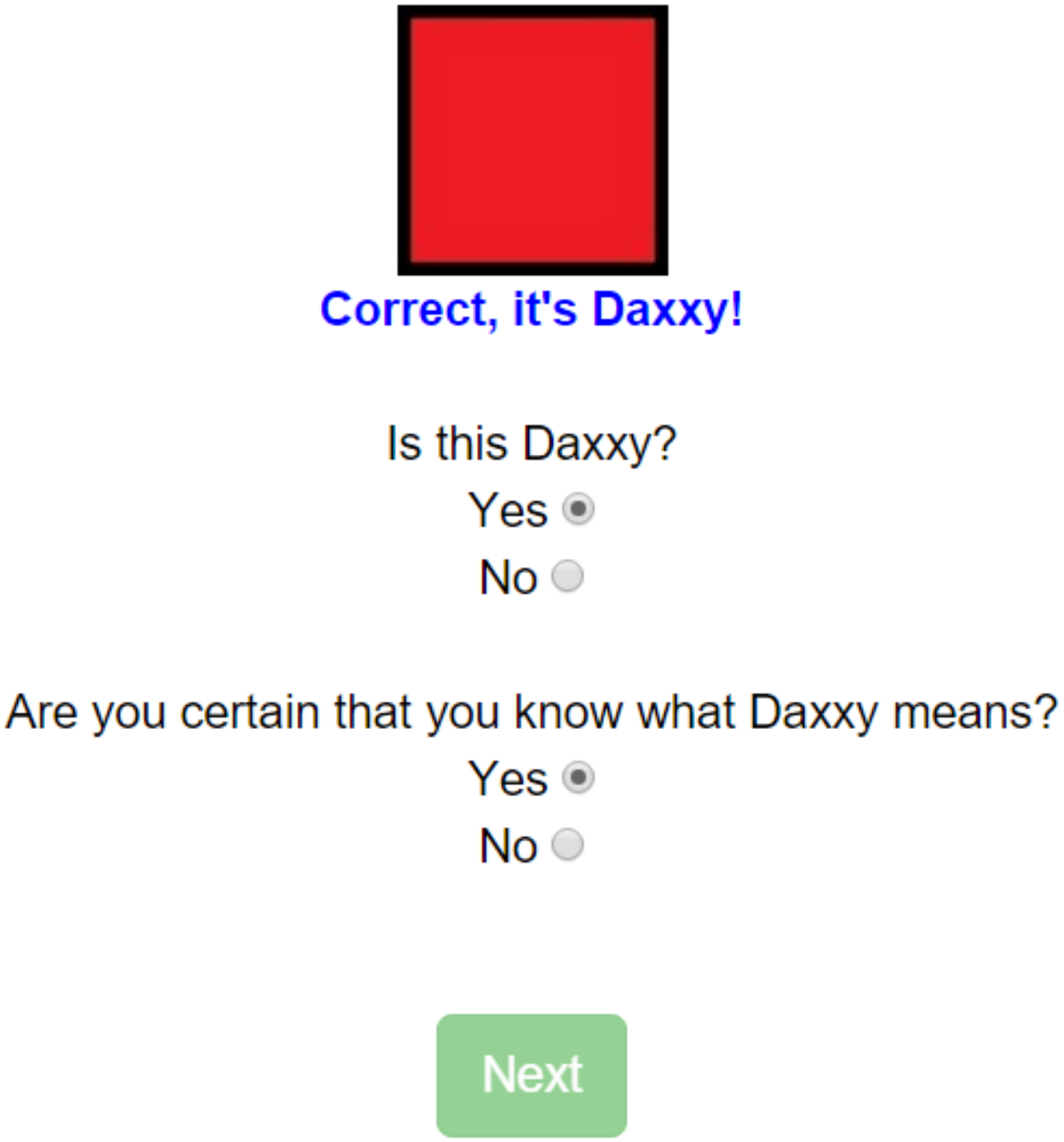


Figure 1. In Experiment 1, participants saw 24 trials (as above), randomized between-conditions. Feedback was displayed after responding.

Results

Table 2. Certainty predictors (behavioral predictors in gray).

Predictor	Description
Trial	Participants become more certain as they complete more of the experiment
Total Accuracy	Total performance thus far
Local Accuracy	Performance on previous N trials ($N = 2, 3, 4, 5$)
Local Accuracy Current	Performance on previous N trials ($N = 2, 3, 4, 5$) and a guess on the current trial
Current Accuracy	Performance on the current trial
Entropy	Uncertainty over hypotheses regarding what the concept is
Domain Entropy	Uncertainty over which objects belong to the concept
Change in Entropy	Entropy change from the previous trial
Change in Domain Entropy	Domain entropy change from the previous trial
Cross Entropy	How much beliefs about hypotheses have changed since the previous trial
Domain Cross Entropy	How much beliefs about which objects belong to the concept have changed since the previous trial
MAP	The probability of the best hypothesis
Maximum Likelihood	The probability of the best hypothesis ignoring the prior probability
Response Probability	Posterior probability of a participant responding they are certain to a particular piece of data

We first visualize plots of participants' certainty and accuracy for each concept in order to show *(i)* whether certainty and accuracy improved over the course of the experiment, *(ii)* whether theoretically harder concepts (according to Feldman, 2000) were, in fact, more difficult for participants, and *(iii)* 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 accuracy curves indicate participants learned the concepts in some conditions but not others. This is beneficial to our analysis as it allows us to analyze conditions and trials in which participants should have had high uncertainty. Overall, participant certainty was inversely proportional to concept difficulty. Participant certainty generally increased, but only reached high values in conditions in which they also achieved high accuracy. The increasing trend of certainty in conditions for which accuracy did not go

above 50% may be reflective of overconfidence. It is also important to note that even though participants received exhaustive evidence, there were still multiple logical rules that were both equivalent and correct. Despite this, participants still became certain over time.

We will first consider our predictors as separate models in order to determine which best predict certainty. Subsequently we will build a model using the best predictors of each type in order to determine the unique contributions of each predictor.

We assessed our predictors with generalized logistic mixed effect models fit by maximum likelihood with random subject and condition effects.[1] First, this analysis shows model accuracy significantly predicts behavioral accuracy (R^2 = .50, β = .748, z = 30.423, p $<$ .001; Figure 3), meaning that overall performance can be reasonably well predicted by the learning model.

Figure 4 then shows mean certainty responses for each trial and condition (y-axis) over several different key predictors of certainty (x-axis). A perfect model here would have data points lying along the line $y = x$ with very little residual variance. **Local Accuracy 5 Back**, the accuracy averaged over the past 5 items, and has low residual variance, meaning that individuals with low local accuracy were uncertain and individuals with high local accuracy were highly certain. Likewise, **Domain Entropy** also has low residual variance and is highly ordered compared to the other model predictors (See Figure A.1 for additional predictor visualizations).

Table A.2 shows the full model results, giving the performance of each model in predicting certainty ratings.[2] These have been sorted by AIC, which quantifies the fit of each model penalizing its number of free parameters (closer to $-\infty$ is better). The AIC score is derived from a generalized logistic mixed effect model fit by maximum likelihood with random subject and condition effects. This table also provides an R^2 measure, calculated using the Pearson correlation between the means of each response and predictor for each trial and condition (this ignores variance from participants). As this table makes clear, the behavioral predictors tend to outperform the model predictors, at times by a substantial amount. The best predictor, **Local Accuracy 5 Back** accounts for 58% of the variance. Additionally, **Local**

[1] We also analyzed our data on an individual level in order to ensure our findings were not due to averaging effects (Estes & Todd Maddox, 2005). See Table A.1 in Supplemental Materials.

[2] See Table A.3 in Supplemental Materials for simplified grammar predictors.

Accuracy models outperform most of the other alternatives, a pattern which is robust to the way in which local accuracy is quantified (e.g., the number back that were counted or whether the current trial is included). The quantitatively best **Local Accuracy** model tracks accuracy over the past five trials. One possible explanation for this is that participants were simply basing their certainty on recent performance. The high performance of both **Local Accuracy** and **Total Correct** implies that people's certainty is largely influenced by their own perception of how well they were doing on the task.

Strikingly, the lackluster performance of the majority of ideal learner models suggests that subjective certainty is not calibrated to the ideal learner. This is consistent with the theory that learners were likely not maintaining more than one hypothesis—perhaps they stored a sample from the posterior, but did 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 current trial correct. Subjective certainty does not accurately predict accuracy on the current trial, or vice versa.

Table 3. Regression for best predictors in Experiment 1 (behavioral predictors in gray).

Predictor	Beta	Standard Error	z value	p
Intercept	-0.82	0.02	-37.61	$< .001$
Local Accuracy 5 Back	0.69	0.04	19.82	$< .001$
Log Trial	-0.60	0.04	-13.93	$< .001$
Total Correct	0.54	0.04	12.00	$< .001$
Domain Entropy	-0.34	0.06	-5.91	$< .001$
Entropy	-0.10	0.05	-1.93	0.054
Log Maximum Likelihood	-0.04	0.04	-1.11	0.269

Our first analysis treated each predictor separately and found the best, but what if multiple predictors were jointly allowed to predict certainty? To answer this, we created a model using the top three behavioral predictors and the top three model-predictors in order to determine the unique contributions of

each (see Table 3).[3][4] As the table makes clear, all behavioral predictors, along with **Domain Entropy**, make significant, unique contributions to certainty. That **Domain Entropy** is significant in this regression but not a leader in AIC simply means that it (or something correlated with it) contributes to certainty judgments despite not being a primary determinant. Conversely, **Entropy** and **Log Maximum Likelihood** were not significant when controlling for the other predictors, demonstrating they provide no unique contributions to certainty. In alignment with the results of our AIC analysis, the (normalized) beta weights, which quantify the strength of each predictors' influence, reveal that the behavioral predictors have the largest influence.

Discussion

Our results showed that an ideal learning model predicts learners' accuracy in our task. These results hold regardless of whether certainty is measured on a binary, or a continuous scale (see Experiment 4 in Supplemental Materials Appendix D). A plausible hypothesis would then be that the predictors derived from our ideal learning model would also be related to learners' certainty, perhaps to a large degree. Instead, we found that **Local Accuracy** and **Total Correct** are most predictive of people's certainty, outperforming our other predictors by predicting as much as 58% of the possible variance. In fact, overwhelmingly, the behavioral predictors performed better than the model predictors. **Domain Entropy** performs well and even has the highest R^2 value, however it is important to emphasize that these R^2 values do not take individual variance or the null model into account. Overall, the results suggested that participants primarily used the feedback on each trial in order to guide their senses of uncertainty about the concept.

EXPERIMENT 2

Motivation

Experiment 1 leaves open the possibility that both **Local Accuracy** and model-based predictors influence behavior, but that feedback overshadowed other predictors, perhaps because feedback was a quick and

[3] This regression was moderately sensitive to which predictors were included, likely due to some degree of multicollinearity.

[4] It was not possible to use random slopes in this regression due to a lack of convergence (Barr, Levy, Scheepers, & Tily, 2013).

reliable cue. Experiment 2 tested this by removing feedback and thus removing it as a cue. We accomplished this by providing participants with only a *single* trial.

The critical question is whether the model-based predictors will become *more* predictive of responses compared to Experiment 1. If so, the cues to certainty may be strategically chosen based on what is informative, with participants able to use model-based measures when information about performance is absent. Alternatively, if the model-based predictors do not improve relative to Experiment 1, that would suggest that factors like **Local Accuracy** may be *the* driving force in metacognitive certainty and absent these predictors, people do not fall back on other systems.

Methods

Like Experiment 1, Experiment 2 presented participants with the task of discovering a hidden Boolean rule (see Figure 5 and Figure 6). We tested 577 participants via Amazon Mechanical Turk on a single-trial version of the same task used in Experiment 1, using the same set of concepts. The experimental trial tested participants on a single concept and displayed all eight images seen in a block of Experiment 1 simultaneously, each labeled with a "*yes*" or "*no*" to indicate whether it was part of the concept (see Figure 5). The participant answered whether they were certain what the concept was. They then saw the same set of eight images (randomized by condition) and were asked to label each as being a part of the concept (see Figure 6). (See Supplemental Materials Appendix B for further detail.)

Results

Figure B.1 shows participants' certainty and accuracy for each condition. Unlike Experiment 1, accuracy was high across most conditions. This was likely due to the ability to view the concept space simultaneously and being tested immediately afterwards. Such a format would make it much easier to determine the concept and lead to reduced memory demands compared to Experiment 1. Despite this, subjective certainty was similar to Experiment 1 in that it related inversely to concept difficulty. Thus, since information regarding the underlying concept was still encoded and used in calculating their certainty, task differences did not seem to influence their certainty.

For Experiment 2, we assessed our predictors with generalized logistic mixed effect models fit by maximum likelihood with random condition effects. Unlike Experiment 1, the model fit for accuracy in

Experiment 2 is not significant (R^2 = .02, β = -.049, z = -1.114, p = .265; Figure B.2). This is likely due to data sparsity, although it is possible that participants did not learn these concepts as well due to the presentation format. In evaluating predictors of certainty Figure B.3 and Table B.1 makes clear that the results are similar to Experiment 1, with the best-performing predictors being behavioral measures.[5] In this case, the only behavioral predictor, **Total Correct** is also the best predictor of certainty. Likewise, while **Domain Entropy** is the best performing model predictor, it is not as good as **Total Correct**. This is strong evidence that removing feedback had little to no effect on participants' propensity to avoid model-based predictors when constructing their own subjective certainty.

Discussion

Our results demonstrate that feedback is not overriding model-based predictors when participants evaluate subjective certainty. When feedback is removed, participants still primarily used a behavioral predictor of overall accuracy in evaluating their own certainty. This could plausibly be because behavioral predictors provide a low-cost and rapid way of calculating certainty while model-based predictors are non-obvious and require more complex calculations.

Experiment 3

Motivation

Both Experiment 1 and Experiment 2 asked about participants' certainty about a target *concept* that was underlying all of the observed data ("Are you certain you know what Daxxy means?"). However, word meanings are highly context dependant. A participant may be highly certain they know the meaning of "Daxxy" within the confines of the experiment but highly uncertain in general. Additionally, other work on metacognition has examined participants' certainty about their current *response*, where model-based effects can sometimes be seen. Experiment 3 examined trial-based certainty measures using the same setup of logical rules used in Experiments 1 & 2. If we find behavioral predictors no longer predict certainty but model-based predictors do, this would provide strong evidence that trial-certainty and concept-certainty are informed by two distinct processes.

[5] See Table B.2 for simplified grammar predictors.

Methods

Experiment 3 was a variant of Experiment 1 in which instead of asking "Are you certain that you know what Daxxy means?" we asked "Are you certain you're right?" after each response (see Supplemental Materials Appendix C). We tested 536 participants on Amazon Mechanical Turk, using otherwise identical methods to Experiment 1 (see Supplemental Materials Appendix C for further details).

Results

Figure C.1 shows participants' certainty and accuracy across trials in each condition for Experiment 3. Unsurprisingly, participant accuracies were similar to Experiment 1, replicating the general observed trends. Importantly, however, certainty in Experiment 3 seems to much more closely track accuracy on each trial, meaning that it is likely veridically reflecting participants' knowledge of each item response (as opposed to the meaning of "daxxy"). We assessed our predictors with generalized logistic mixed effect models fit by maximum likelihood with random subject and condition effects. Like Experiment 1, the model fit between behavioral and model accuracy in Experiment 3 is reliable, (R^2 = .50, β = .808, z = 31.529, p $<$ .001; Figure C.2).

Figure C.3 shows subjective certainty (y-axis) over many key predictors (x-axis). Again, a perfect model would have data points lying along the line $y = x$ with very little residual variance. Once again, **Local Accuracy** predictors trend in this direction and have low residual variance. Model-based predictors look similar to Experiment 1, with many having large amounts of residual variance.

Table C.1 shows the full model results for Experiment 3, sorted by AIC and giving the performance of each model in predicting certainty ratings.[6] Behavioral predictors once again overwhelmingly outperform the model-based predictors. Similar to Experiment 1, **Local Accuracy 5 Back Current** is the best predictor at 70% of variance explained, and the best model-based predictor is again **Domain Entropy** which accounts for 61% of the variance.

Discussion

[6] See Table C.2 for simplified grammar predictors.

Experiment 3 provides strong evidence that participants primarily relied on local accuracy for their trial-based certainty just as they did for concept-based certainty. This reflects the fact that trial-based certainty, while more independent than concept-based certainty per trial, was still influenced by performance and feedback on previous trials. Like Experiment 1, participants did not seem to be using most model-based predictors in their certainty calculations, despite behaving in-line with model predictions with regard to accuracy. These results are seemingly in conflict with the Sanders et al. (2016) model which they demonstrated to be a good predictor of participant certainty. One possibility is that these differences were the result of cross-trial learning in our task required. Neither Sanders et al. (2016) tasks required such cross-trial learning.

GENERAL DISCUSSION

In conjunction with past research, our results paint a picture of how subjective certainty is derived for high-level logical domains like Boolean concept learning. It appears that certainty estimation primarily makes use of behavioral and overt task features, but that some model-predictors are also relevant. In contrast, certainty about previous learning and perceptual certainty seem to default to using predictors derived from ideal learning models.

In Experiments 1 and 3, **Local Accuracy** and **Total Correct** were very successful predictors of certainty. This means that participants seemed to primarily be basing their certainty on their past performance—*inferring certainty from their own behavior and feedback.* If certainty was fulfilling its function, one might expect **Current Accuracy** to be an excellent predictor. Instead, we find it is an extremely poor predictor, implying that people's sense of certainty in these tasks is not calibrated well to their future performance. This is also in line with past research showing that some people's certainty is not based solely on their perceived probability of being correct, but also on the inverse variance of the data (Navajas et al., 2017). This general pattern is not unlike findings from metacognitive studies showing that often people do not understand—or perhaps even remember—the causes of their own behavior (Johansson, Hall, Sikström, & Olsson, 2005; Nisbett & Wilson, 1977). People do not 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 when feedback is present and learning is taking place. In these cases, people do not appear to reflect an awareness of how much certainty they *should* have.

Past studies in memory have found that initial eyewitness confidence reliably predict eyewitness accuracy but confidence judgments after memory "contamination" has occurred were no longer reliable (Wixted, Mickes, Clark, Gronlund, & Roediger III, 2015). Given our results, a possible explanation for this is that the feedback in our experiments played the same role as the memory contamination in the eyewitness studies. In other words, recent feedback heavily influences certainty, and if that feedback is unreliable, it could lead to false memories.

It should be noted that one possible reason the behavioral predictors outperform the model predictors is that the behavioral predictors will vary with participants' mental states and thus with the natural idiosyncrasies within, although this effect may be mitigated by our used of mixed-effect models. For example, individual differences in attention that influence performance a by-subject level could be captured by the behavioral predictors, but not the model-based predictors which are functions only of the observed data. Though difficult to quantitatively evaluate, this difference may in part explain why the behavioral predictors are dominant in capturing performance, and this possible mechanism is consistent with the idea that certainty is primarily derived from observing our own behavior and secondarily by the properties of the data.

Our analyses also help inform us about which factors *do not* drive certainty during learning, and several are surprising. One reasonable theory posits that participants could base their certainty off of their confidence in the Maximum a Posteriori (MAP) hypothesis under consideration. Since the MAP predictors do not perform well, it is unlikely that learners' certainty relies on internal estimates of the probabilities of the most likely hypothesis.

CONCLUSION

Our findings suggest that although several types of predictors make unique contributions to certainty, the primary predictors of certainty are from observations of people's own behavior and performance, not from measures derived from an idealized learning model. Although learning patterns follow an idealized mathematical model, subjective certainty is only secondarily influenced by that model regardless of whether or not they were able to observe how well they were doing. This is likely due to the underlying process of hypothesis formation and revision, as well as the way in which probabilities are handled beyond that which an ideal learner provides. These results also provide counterintuitive insight into why

humans become certain. Certainty about a latent, abstract concept does not seem to be determined by the same mechanisms that drive learning, and a large component of certainty could reflect factors that are largely removed from the veridical probabilities that any given hypothesis is correct.

ACKNOWLEDGMENTS

We thank the Jacobs Foundation, the Google Faculty Research Awards Program, and the National Science Foundation Research Traineeship Program (grant number 1449828) for the funding to complete this work. We also thank members of the Kidd Lab and the Computation and Language Lab for providing valuable feedback.

REFERENCES

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*, 149–178.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

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

Bruner, J. S., & Austin, G. A. (1986). *A study of thinking*. Transaction publishers.

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between belief and performance in perceptual decision making. *PloS one*, *9*(5), e96511.

Estes, W. K., & Todd Maddox, W. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*(3), 403–408.

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.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, *24*(3), 411–435.

Halpern, J. Y., & Fagin, R. (1992). Two views of belief: belief as generalized probability and belief as evidence. *Artificial intelligence*, *54*(3), 275–317.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the*

National Academy of Sciences, *112*(33), 10321–10324.

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.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

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.

Loftus, E. F., Donders, K., Hoffman, H. G., & Schooler, J. W. (1989). Creating new memories that are quickly accessed and confidently held. *Memory & Cognition*, *17*(5), 607–616.

Marks, G., & Miller, N. (1985). The effect of certainty on consensus judgments. *Personality and Social Psychology Bulletin*, *11*(2), 165–177.

McDermott, K. B., & Roediger, H. L. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language*, *39*(3), 508–520.

Miller, S. A. (1986). Certainty and necessity in the understanding of piagetian concepts. *Developmental Psychology*, *22*(1), 3.

Miller, S. A., Brownell, C. A., & Zukier, H. (1977). Cognitive certainty in children: Effects of concept, developmental level, and method of assessment. *Developmental Psychology*, *13*(3), 236.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *bioRxiv*, 102269.

Nelson, T., Narens, L., & Bower, G. (1990). The psychology of learning and motivation. *Metamemory: A theoretical framework and new findings*.

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.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535.

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence.

Neuron, *90*(3), 499–506.

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. (2002). What doesn't kill me makes me stronger: the effects of resisting persuasion on attitude certainty. *Journal of personality and social psychology*, *83*(6), 1298.

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.

Tsai, C. I., Klayman, J., & Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. *Organizational Behavior and Human Decision Processes*, *107*(2), 97–105.

van der Lubbe, J. C., Boxma, Y., & Boekee, D. E. (1984). A generalized class of certainty and information measures. *Information Sciences*, *32*(3), 187–215.

Wells, G. L., & Bradfield, A. L. (1998). "good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*(3), 360.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger III, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*(6), 515.

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: why might people be more confident in their less accurate judgments? *Journal of experimental psychology. Learning, memory, and cognition*, *35*(2), 558.

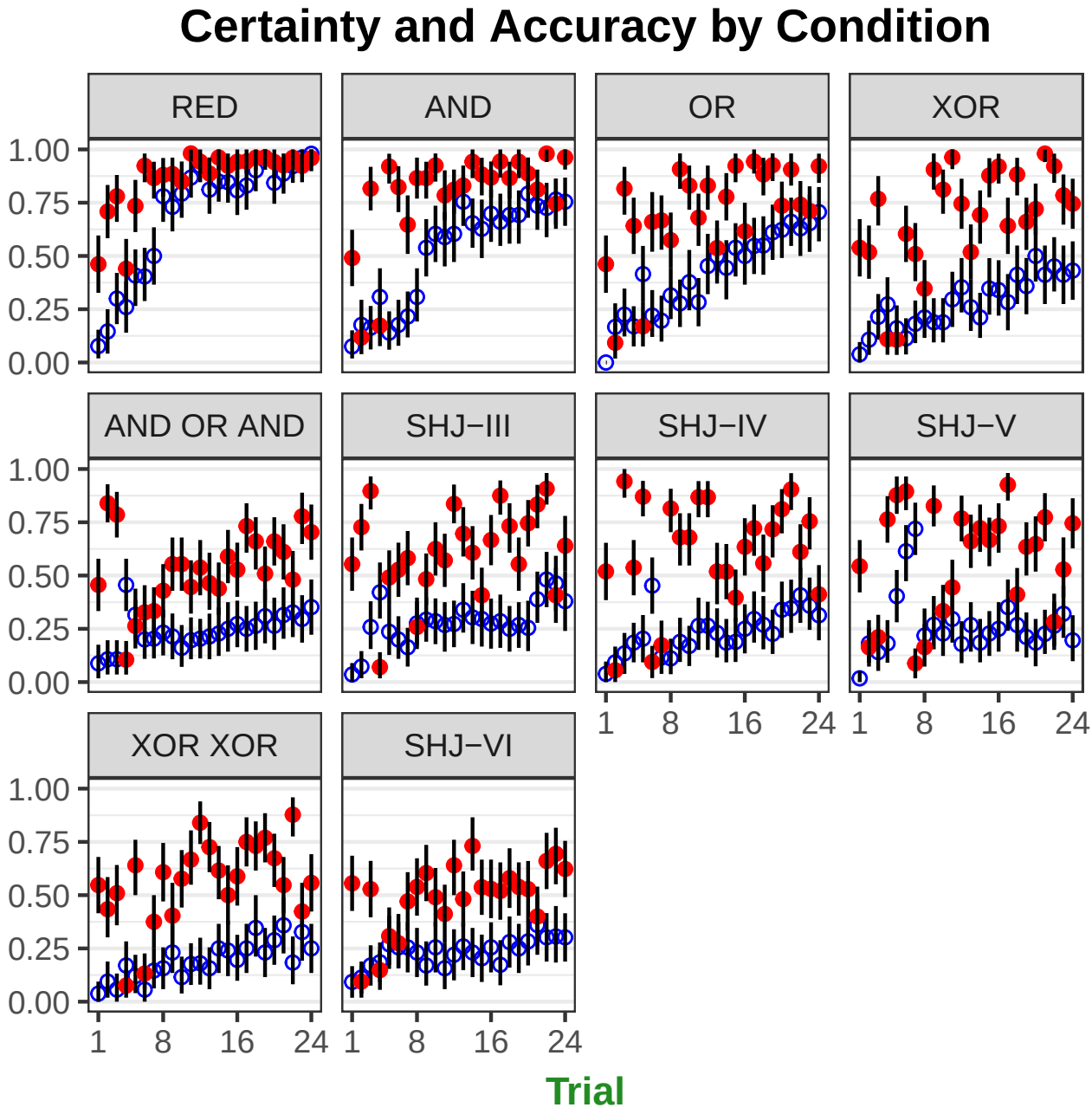
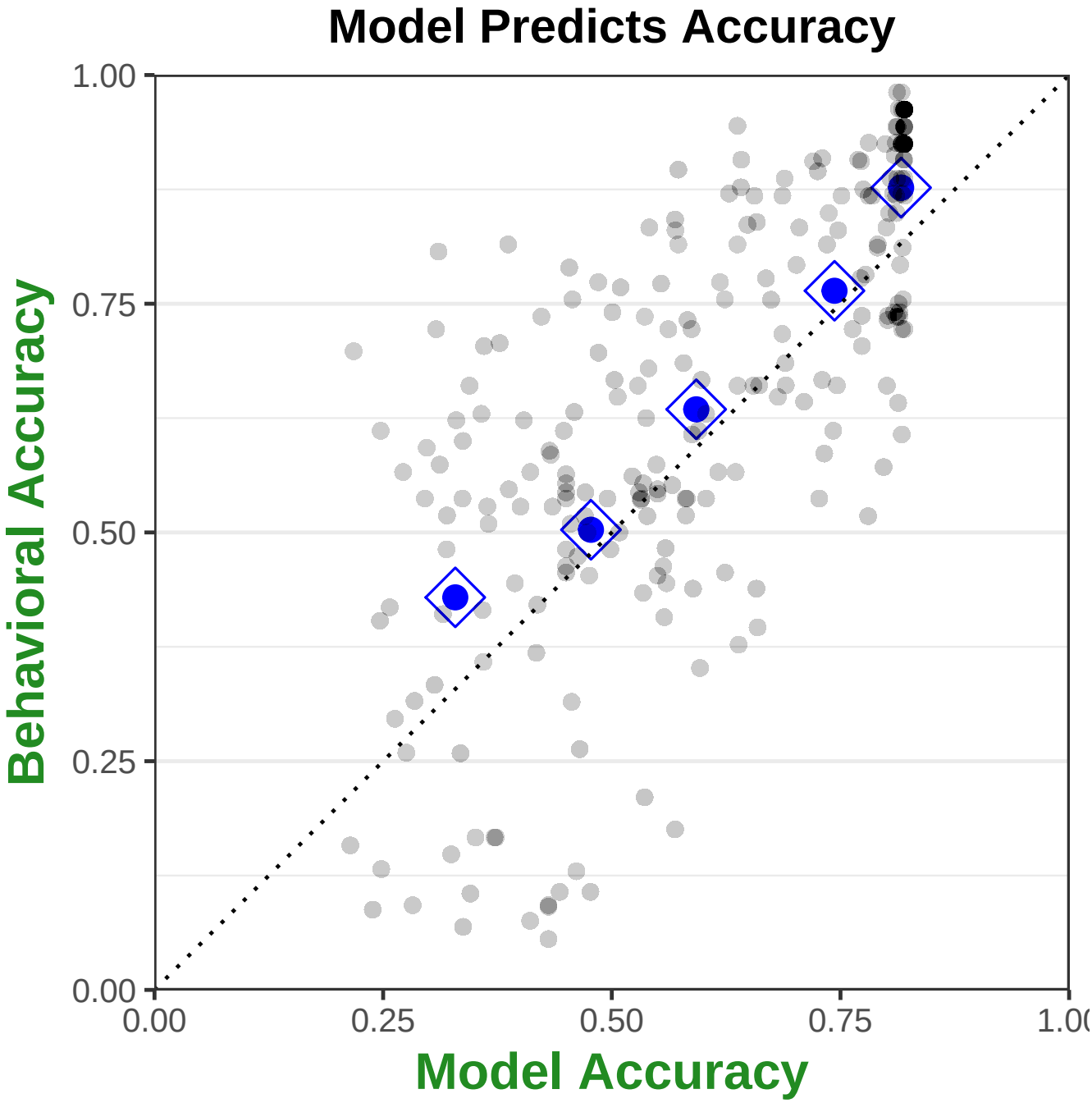


Figure 2. Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 1. Chance is 50% across all conditions if guesses are made randomly.



116 **Figure 3.** Model vs. behavioral accuracy for Experiment 1

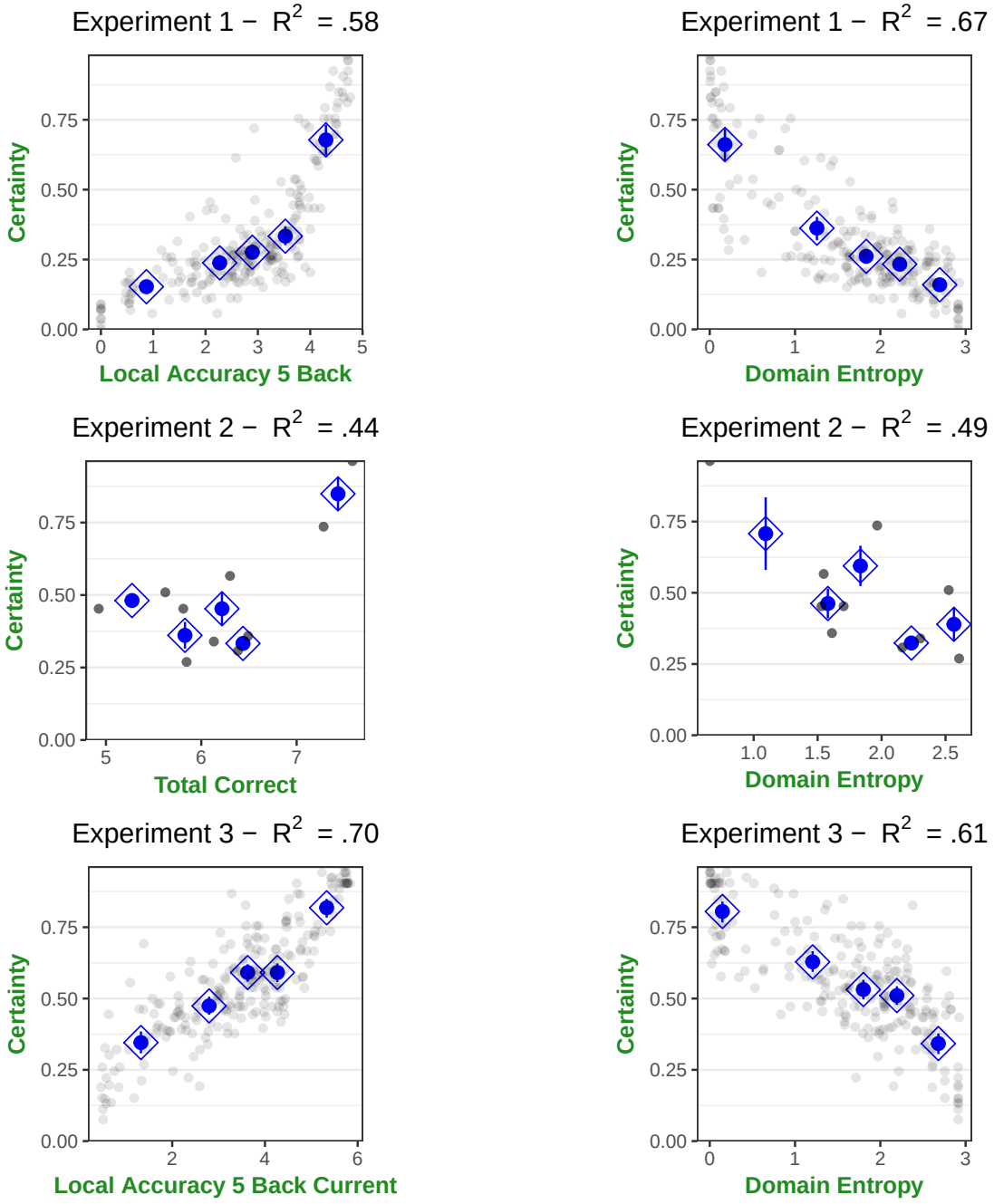
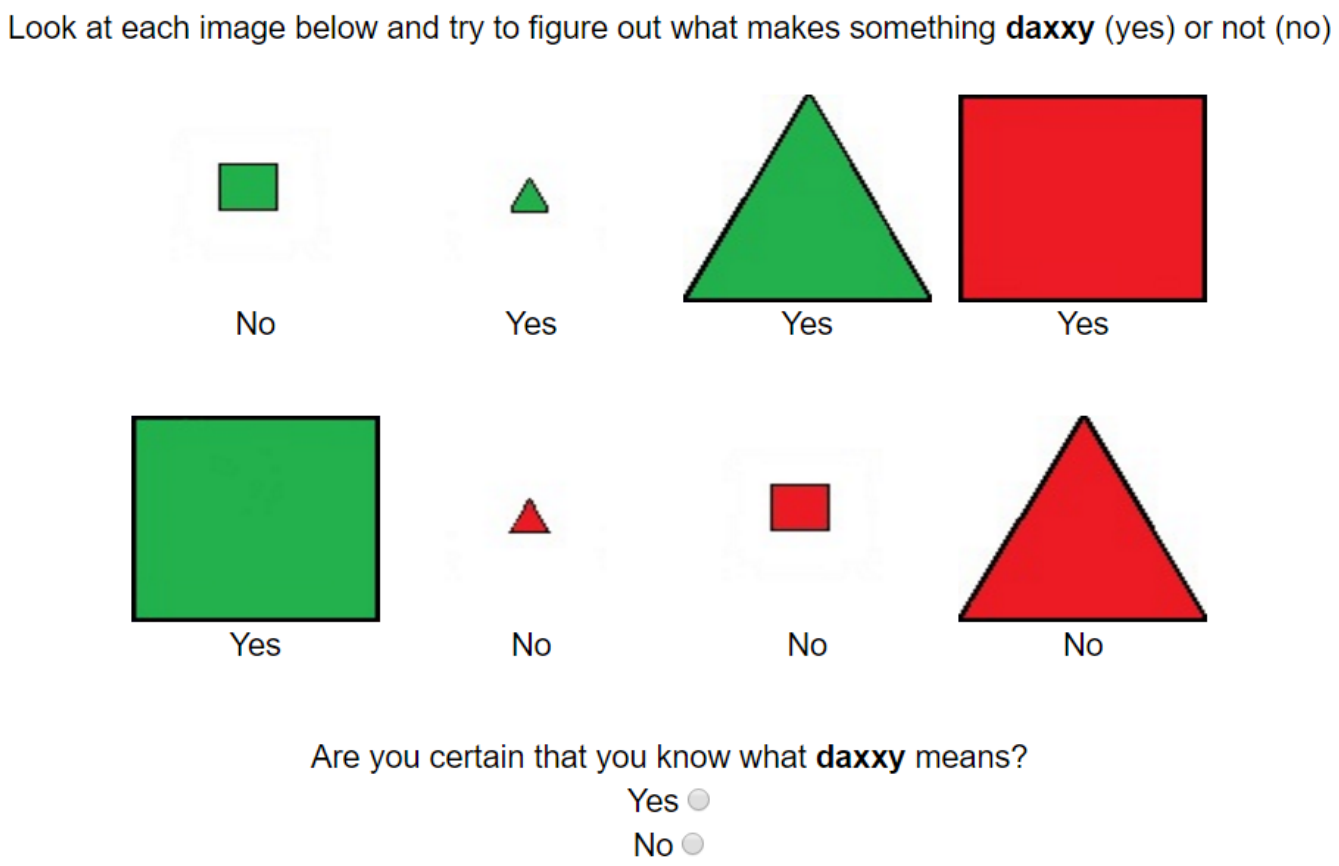
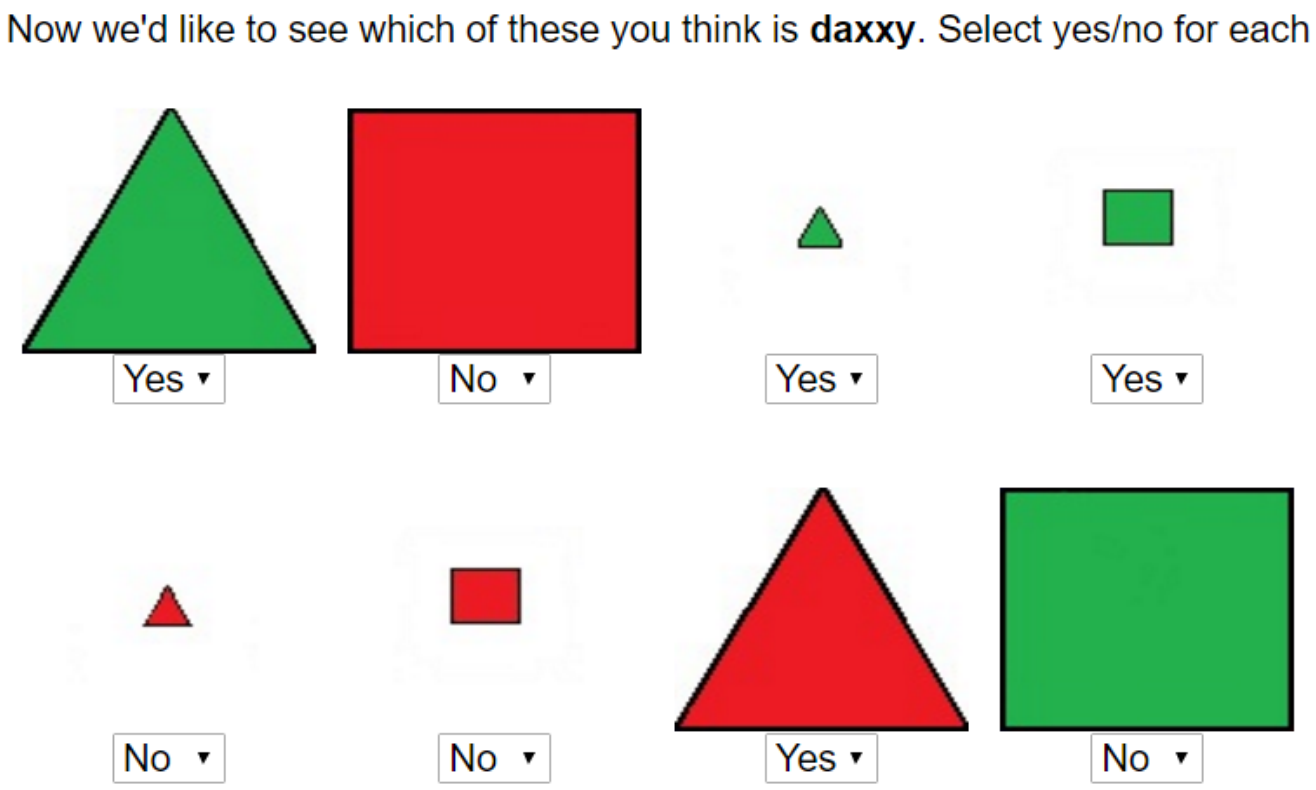


Figure 4. Key model fits for Experiments 1-3, showing mean participant responses for each concept and trial (gray) and binned model means in each of five quantiles (blue) for certainty rating (y-axis) as a function of model (x-axis). Diagonal lines with low variance correspond to models which accurately capture human behavior.



191 **Figure 5.** In Experiment 2, participants saw a single trial (as above), randomized between-conditions.



192 **Figure 6.** In Experiment 2, after responding regarding their certainty, participants labeled each stimuli to assess their accuracy.

SUPPLEMENTAL MATERIALS

A: EXPERIMENT 1 METHODS

552 participants were recruited 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 discern the meaning of a word that represented a specific concept. Participants practiced on eight practice trials to ensure that they understood the task before proceeding to the actual study. For the experimental trials (see Figure 1), participants saw one of ten conditions, each composed of 24 trials. During each trial, participants guessed whether the object fit the undisclosed concept by responding "*yes*" or "*no*". Participants also reported whether or not they were certain about the meaning of the novel word.[7] At the end of each trial, participants received correct/incorrect feedback about their guess.

Each condition represented one unique concept of varying complexity (see Table 1), such that each participant made judgments for only one concept. Following the Shepard et al. (1961) experiment, stimuli spanned three binary dimensions: shape (square or triangle), color (red or green), and size (large or small). Regardless of condition, participants saw the same set of eight images (which exhaustively spanned the space) in blocks of three. The ordering of the images was randomized between-conditions.

Concepts 1, and 5-9 (Table 1) are identical to concepts used in both the Shepard et al. (1961) and Feldman (2000) experiments. These concepts spanned the concept family consisting of three features and four positive examples. Additional conditions were added to test for potential differences between operators.

In order to address whether learners felt as certain as is justified by the data, we used an ideal learning model to determine how confident a learner should have been. Goodman et al. (2008) used a similar model to formalize concept learning in a probabilistic setting, in which notions of certainty and uncertainty (e.g. Shannon, 1948) were well defined. Our implementation was developed using Python and the Language Of Thought library, *LOTlib* (Piantadosi, 2014). The model defines a probabilistic context-free grammar (PCFG) with a set of primitives: red, green, triangle, square, large, small, and logical operations (shown in Table A.4).[8] The PCFG serves as a prior over hypotheses and specifies an infinite hypothesis space. This prior is uniform over each basic rule in the grammar. Due to the

[7] We also ran a version of Experiment 1 which measured certainty on a continuous scale. These results followed the same pattern. See Appendix D.

[8] In order to test additional model-based predictors we also ran our model using a simplified grammar. See Table A.5.

multiplication of compositional rules, a simplicity prior arises as more complex rules have lower probability.

While PCFG models might limit the inferences and generalizations we are able to make regarding human cognition, they are also the current state of the art when it comes to predicting accuracy in terms of concept learning. Since the prediction of human accuracy by the model is an essential prerequisite to evaluating the performance of the models in predicting human certainty, PCFG models are the best candidates. In other words, although other models may be better at predicting certainty, they will likely be worse at predicting accuracy and thus, extremely limited in their inferences about human cognition.

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 for each trial. The model incorporated parameters for the noise in the data (alpha) and a power law memory decay on the likelihood of previous data[9] (beta), best fit (on participant accuracy using a grid search) as 0.64 and 0 respectively.

Additionally, logarithmic transformations are common in psychophysics (Stevens, 1957) and therefore, many of our predictors were considered in their standard form, as well as under a logarithmic transformation, yielding a total of 38 models. Some predictors used a $log(1 + x)$ transformation to avoid problems with zeroes.

B: EXPERIMENT 2 METHODS

577 participants on Amazon Mechanical Turk went through two practice trials before the experimental trial. The experimental trial tested participants on a single concept and displayed all eight images seen in a block of Experiment 1. Each image was labelled with a "*yes*" or "*no*" to indicate whether it was part of the concept (see Figure 5). The participant answered whether they were certain what the concept was. They then saw the same set of eight images (randomized by condition) and were asked to label each as being a part of the concept (see Figure 6). Like Experiment 1, our model incorporated noise (alpha) and memory decay (beta) parameters, best fit as 0.65 and 0.06 respectively.

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

C: EXPERIMENT 3 METHODS

536 participants on Amazon Mechanical Turk practiced on eight practice trials to ensure that they understood the task before proceeding to the actual study. For the experimental trials, participants saw one of ten conditions, each composed of of 24 trials. Each condition tested for a different concept with varying complexity (see Table 1).

Like Experiment 1 and 2, our model incorporated parameters for the noise in the data (alpha) and a power law memory decay on the likelihood of previous data (beta), best fit as 0.66 and 0 respectively.

D: EXPERIMENT 4

Motivation

Experiments 1-3 used a binary certainty judgement. In order to test whether our model predictors were failing due to finer certainty gradations being collapsed in our data, we ran a fourth experiment which used a continuous certainty scale.

Methods

Experiment 4 was a variant of Experiment 1 in which instead of asking "Are you certain that you know what Daxxy means?" we asked "How certain are you that you know what Daxxy means?". Participants selected their certainty on a one to five scale with one labelled as "Not at all certain" and 5 labelled as "Very certain". 535 participants on Amazon Mechanical Turk practiced on eight practice trials to ensure that they understood the task before proceeding to the actual study. For the experimental trials, participants saw one of ten conditions, each composed of of 24 trials. Each condition tested for a different concept with varying complexity (see Table 1).

Our model incorporated parameters for the noise in the data (alpha) and a power law memory decay on the likelihood of previous data (beta), best fit as 0.66 and 0 respectively.

Results

Figure D.1 shows participants' certainty and accuracy across trials in each condition for Experiment 4. Unsurprisingly, participant accuracies were similar to Experiment 1 and 3. We also examined the relationship between the continuous certainty scores in Experiment 4 and the binary certainty scores in

Experiment 1 (see Figure D.2) and found that the continuous certainty scores strongly predict the binary scores (R^2 = .93, β = 4.575, z = 6.556, p $<$ .001).

For Experiment 4, we assessed our predictors with linear mixed effect models fit by maximum likelihood with random subject and condition effects. The model fit for accuracy in Experiment 4 is significant, (R^2 = .13, β = .170, t = 36.18, p $<$ .001; Figure D.3).

Figure D.4 shows certainty (y-axis) over many key predictors of certainty (x-axis). Again, a perfect model would have data points lying along the line $y = x$ with very little residual variance. Once again, **Local Accuracy** predictors trend in this direction and have low residual variance. Model-based predictors look similar to Experiment 1, with many having large amounts of residual variance.

Table D.1 shows the full model results for Experiment 4, sorted by AIC and giving the performance of each model in predicting certainty ratings.[10] Behavioral predictors once again overwhelmingly outperform the model-based predictors. Similar to Experiment 1, **Local Accuracy 5 Back Current** is the best predictor at 77% of variance explained, and the best model-based predictor is **Domain Entropy** which accounts for 69% of the variance.

Discussion

Experiment 4 provides evidence that using either a binary or continuous scale of certainty does not impact the performance of the predictors. Using a continuous scale, behavioral predictors still outperformed model-based predictors.

[10] See Table D.2 for simplified grammar predictors.

Table A.1. Predictors of certainty rankings for DNF grammar in Experiment 1 when analyzing data by participant vs. as a group. (behavioral predictors in gray).

Ranking	Individual Analysis	Group Analysis
1	Total Correct	Local Accuracy 5 Back
2	Trial	Local Accuracy 4 Back
3	Log Total Correct	Local Accuracy 5 Back Current
4	Log Trial	Domain Entropy
5	Log Local Accuracy 4 Back	Log Local Accuracy 5 Back
6	Log Local Accuracy 5 Back	Total Correct
7	Domain Entropy	Local Accuracy 4 Back Current
8	Local Accuracy 5 Back	Local Accuracy 3 Back
9	Local Accuracy 4 Back	Log Local Accuracy 4 Back
10	Entropy	Log Total Correct
11	Log Local Accuracy 3 Back	Entropy
12	Local Accuracy 5 Back Current	Log Local Accuracy 5 Back Current
13	Log Local Accuracy 5 Back Current	Log Local Accuracy 3 Back
14	Local Accuracy 3 Back	Local Accuracy 3 Back Current
15	Log Max Likelihood	Log Local Accuracy 4 Back Current
16	Log Local Accuracy 4 Back Current	Log Max Likelihood
17	Max Likelihood	Log Trial
18	Log Local Accuracy 3 Back Current	Local Accuracy 2 Back
19	Local Accuracy 4 Back Current	Log Local Accuracy 3 Back Current
20	Local Accuracy 3 Back Current	Trial
21	Local Accuracy 2 Back	Log Local Accuracy 2 Back
22	Log Local Accuracy 2 Back	Local Accuracy 2 Back Current
23	Log Local Accuracy 2 Back Current	Log Local Accuracy 2 Back Current
24	Local Accuracy 2 Back Current	MAP
25	Log MAP	Max Likelihood
26	MAP	Local Accuracy 1 Back
27	Log Local Accuracy 1 Back	Log Local Accuracy 1 Back

386 **Table A.2.** Predictors of certainty for Experiment 1 (behavioral predictors in gray).

Model	AIC	R^2	Log Likelihood	Beta	Standard Error	p
Local Accuracy 5 Back	9644.2	0.58	-4818.1	1.30	0.04	$< .001$
Local Accuracy 4 Back	9735.4	0.59	-4863.7	1.27	0.04	$< .001$
Local Accuracy 5 Back Current	9785.4	0.60	-4888.7	1.26	0.04	$< .001$
Domain Entropy	9799.1	0.67	-4895.5	-1.47	0.04	$< .001$
Log Local Accuracy 5 Back	9851.5	0.44	-4921.8	1.27	0.04	$< .001$
Total Correct	9873.8	0.45	-4932.9	1.13	0.03	$< .001$
Local Accuracy 4 Back Current	9900.8	0.61	-4946.4	1.22	0.04	$< .001$
Local Accuracy 3 Back	9915.2	0.59	-4953.6	1.18	0.03	$< .001$
Log Local Accuracy 4 Back	9920.7	0.46	-4956.4	1.24	0.04	$< .001$
Log Total Correct	9963.3	0.39	-4977.6	1.12	0.03	$< .001$
Entropy	9973.8	0.55	-4982.9	-1.44	0.04	$< .001$
Log Local Accuracy 5 Back Current	10010.1	0.47	-5001.0	1.22	0.04	$< .001$
Log Local Accuracy 3 Back	10072.8	0.48	-5032.4	1.15	0.04	$< .001$
Local Accuracy 3 Back Current	10093.2	0.62	-5042.6	1.13	0.03	$< .001$
Log Local Accuracy 4 Back Current	10099.9	0.49	-5045.9	1.18	0.04	$< .001$
Log Max Likelihood	10102.0	0.35	-5047.0	1.33	0.04	$< .001$
Log Trial	10187.1	0.24	-5089.6	1.01	0.03	$< .001$
Local Accuracy 2 Back	10248.5	0.56	-5120.3	1.00	0.03	$< .001$
Log Local Accuracy 3 Back Current	10266.1	0.51	-5129.1	1.10	0.04	$< .001$
Trial	10338.9	0.22	-5165.4	0.88	0.03	$< .001$
Log Local Accuracy 2 Back	10360.7	0.48	-5176.4	0.98	0.03	$< .001$
Local Accuracy 2 Back Current	10449.4	0.59	-5220.7	0.93	0.03	$< .001$
Log Local Accuracy 2 Back Current	10571.3	0.51	-5281.7	0.90	0.03	$< .001$
MAP	10689.2	0.37	-5340.6	0.96	0.04	$< .001$
Max Likelihood	10694.1	0.15	-5343.0	1.31	0.06	$< .001$
Local Accuracy 1 Back	10787.7	0.42	-5389.8	0.69	0.03	$< .001$
Log Local Accuracy 1 Back	10787.7	0.38	-5389.8	0.69	0.03	$< .001$

Table A.3. Predictors of certainty for Experiment 1 using simplified grammar (behavioral predictors in gray).

Model	AIC	R^2	Log Likelihood	Beta	Standard Error	p
Local Accuracy 5 Back	9644.2	0.58	-4818.1	1.30	0.04	$< .001$
Local Accuracy 4 Back	9735.4	0.59	-4863.7	1.27	0.04	$< .001$
Local Accuracy 5 Back Current	9785.4	0.60	-4888.7	1.26	0.04	$< .001$
Log Local Accuracy 5 Back	9851.5	0.44	-4921.8	1.27	0.04	$< .001$
Total Correct	9873.8	0.45	-4932.9	1.13	0.03	$< .001$
Local Accuracy 4 Back Current	9900.8	0.61	-4946.4	1.22	0.04	$< .001$
Local Accuracy 3 Back	9915.2	0.59	-4953.6	1.18	0.03	$< .001$
Log Local Accuracy 4 Back	9920.7	0.46	-4956.4	1.24	0.04	$< .001$
Log Total Correct	9963.3	0.39	-4977.6	1.12	0.03	$< .001$
Log Local Accuracy 5 Back Current	10010.1	0.47	-5001.0	1.22	0.04	$< .001$
Log Local Accuracy 3 Back	10072.8	0.48	-5032.4	1.15	0.04	$< .001$
Local Accuracy 3 Back Current	10093.2	0.62	-5042.6	1.13	0.03	$< .001$
Log Local Accuracy 4 Back Current	10099.9	0.49	-5045.9	1.18	0.04	$< .001$
Log Trial	10187.1	0.24	-5089.6	1.01	0.03	$< .001$
Local Accuracy 2 Back	10248.5	0.56	-5120.3	1.00	0.03	$< .001$
Log Local Accuracy 3 Back Current	10266.1	0.51	-5129.1	1.10	0.04	$< .001$
Trial	10338.9	0.22	-5165.4	0.88	0.03	$< .001$
Log Local Accuracy 2 Back	10360.7	0.48	-5176.4	0.98	0.03	$< .001$
Local Accuracy 2 Back Current	10449.4	0.59	-5220.7	0.93	0.03	$< .001$
Log Max Likelihood	10475.0	0.15	-5233.5	0.88	0.03	$< .001$
Log Local Accuracy 2 Back Current	10571.3	0.51	-5281.7	0.90	0.03	$< .001$
Domain Entropy	10574.3	0.42	-5283.2	-0.91	0.03	$< .001$
Local Accuracy 1 Back	10787.7	0.42	-5389.8	0.69	0.03	$< .001$
Log Local Accuracy 1 Back	10787.7	0.38	-5389.8	0.69	0.03	$< .001$
Local Accuracy 1 Back Current	10925.5	0.48	-5458.7	0.62	0.03	$< .001$
Log Local Accuracy 1 Back Current	10968.0	0.43	-5480.0	0.61	0.03	$< .001$
Max Likelihood	11162.5	0.01	-5577.2	0.42	0.03	$< .001$

388 **Table A.4.** Disjunctive normal form grammar used to generate logical rules in the idealized learning model. The variable x is the current object.

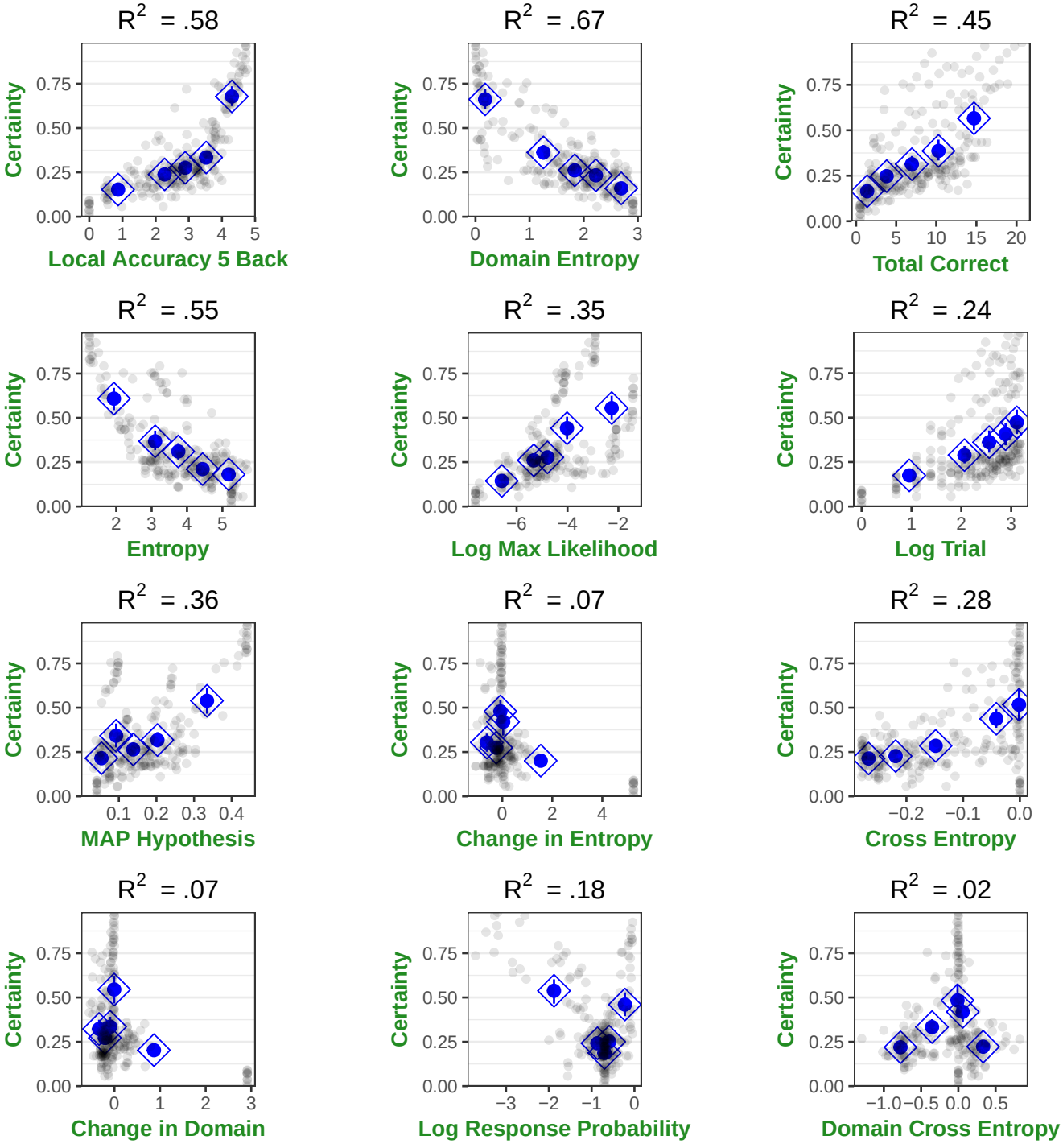
Rule
START → DISJ
DISJ → CONJ
DISJ → or(CONJ, DISJ)
CONJ → BOOL
CONJ → and(BOOL, CONJ)
BOOL → PREDICATE
BOOL → not(PREDICATE)
PREDICATE → red(x)
PREDICATE → green(x)
PREDICATE → triangle(x)
PREDICATE → square(x)
PREDICATE → large(x)
PREDICATE → small(x)

389 **Table A.5.** Simplified grammar

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)

391 **Table B.1.** Predictors of certainty for Experiment 2 (behavioral predictors in gray).

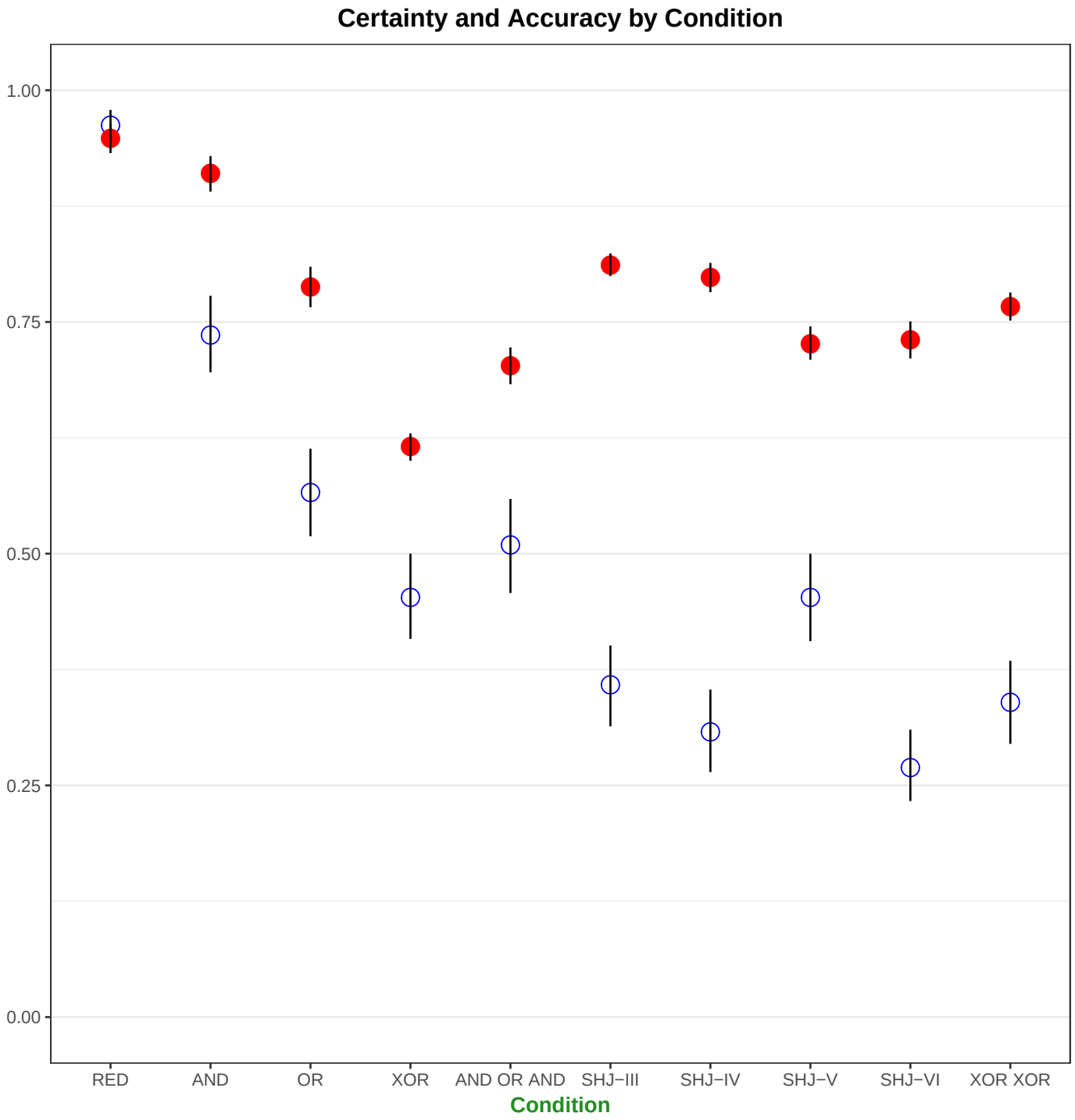
Model	AIC	R^2	Log.Likelihood	Beta	Standard.Error	p
Total Correct	5088.5	0.44	-2541.3	0.12	0.02	$< .001$
Log Total Correct	5100.2	0.39	-2547.1	0.59	0.13	$< .001$
Domain Entropy	5111.7	0.49	-2552.8	-1.58	0.45	$< .001$
Entropy	5115.9	0.24	-2554.9	-0.89	0.42	0.035
MAP	5116.6	0.19	-2555.3	5.89	2.93	0.044
Log Maximum Likelihood	5117.0	0.32	-2555.5	0.60	0.34	0.075
Null Model	5117.9	0.00	-2556.9	-	0.37	0.778
Log MAP	5118.6	0.07	-2556.3	0.64	0.54	0.243
Maximum Likelihood	5118.7	0.15	-2556.4	23.62	6.95	0.001



390 **Figure A.1.** Key model fits for Experiment 1.

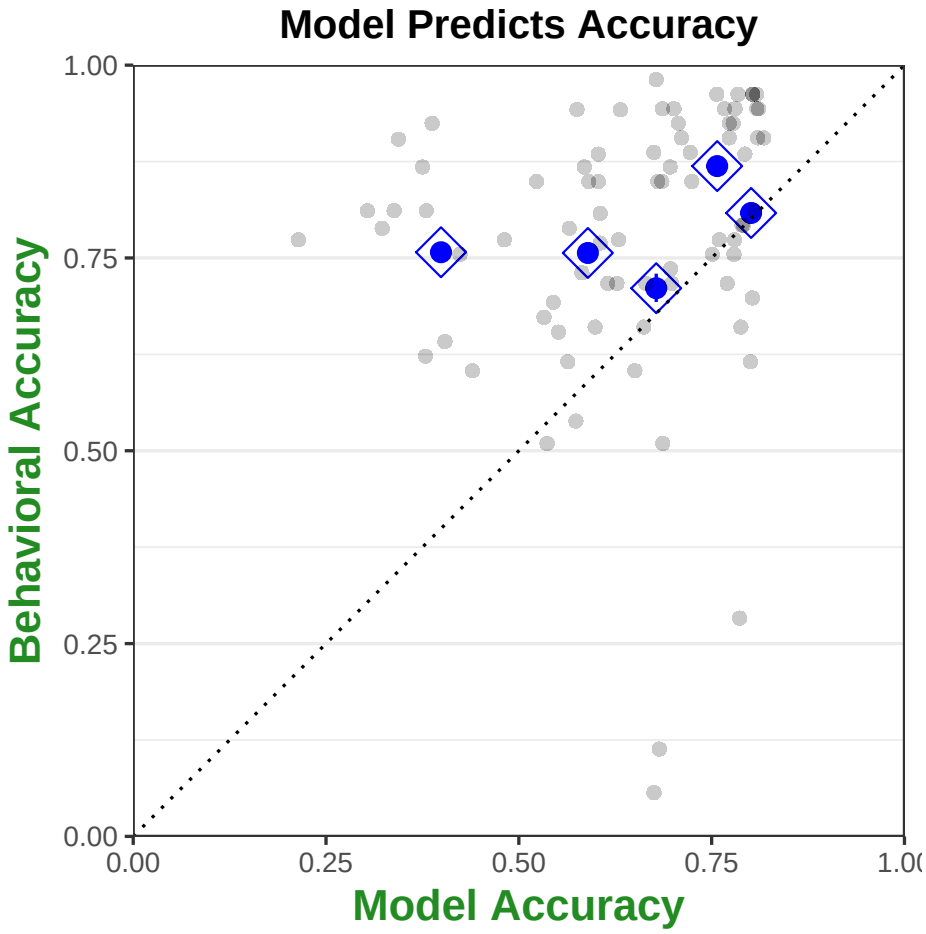
392 **Table B.2.** Predictors of certainty for Experiment 2 using simplified grammar (behavioral predictors in gray).

Model	AIC	R^2	Log.Likelihood	Beta	Standard.Error	p
Total Correct	5088.5	0.44	-2541.3	0.12	0.02	$< .001$
Log Total Correct	5100.2	0.39	-2547.1	0.59	0.13	$< .001$
Log MAP	5116.3	0.35	-2555.1	2.23	1.09	0.041
MAP	5116.8	0.31	-2555.4	3.65	1.94	0.061
Entropy	5117.3	0.26	-2555.7	-1.18	0.68	0.085
Domain Entropy	5117.4	0.22	-2555.7	-0.85	0.50	0.090
Null Model	5117.9	0.00	-2556.9	-	0.37	0.778
Log Maximum Likelihood	5118.5	0.20	-2556.3	0.87	0.74	0.240
Maximum Likelihood	5119.3	0.11	-2556.6	33.91	5.15	$< .001$



393 **Figure B.1.** Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 2. Chance is 50% across all conditions if guesses are

394 made randomly.



395 **Figure B.2.** Model vs. behavioral accuracy for Experiment 2

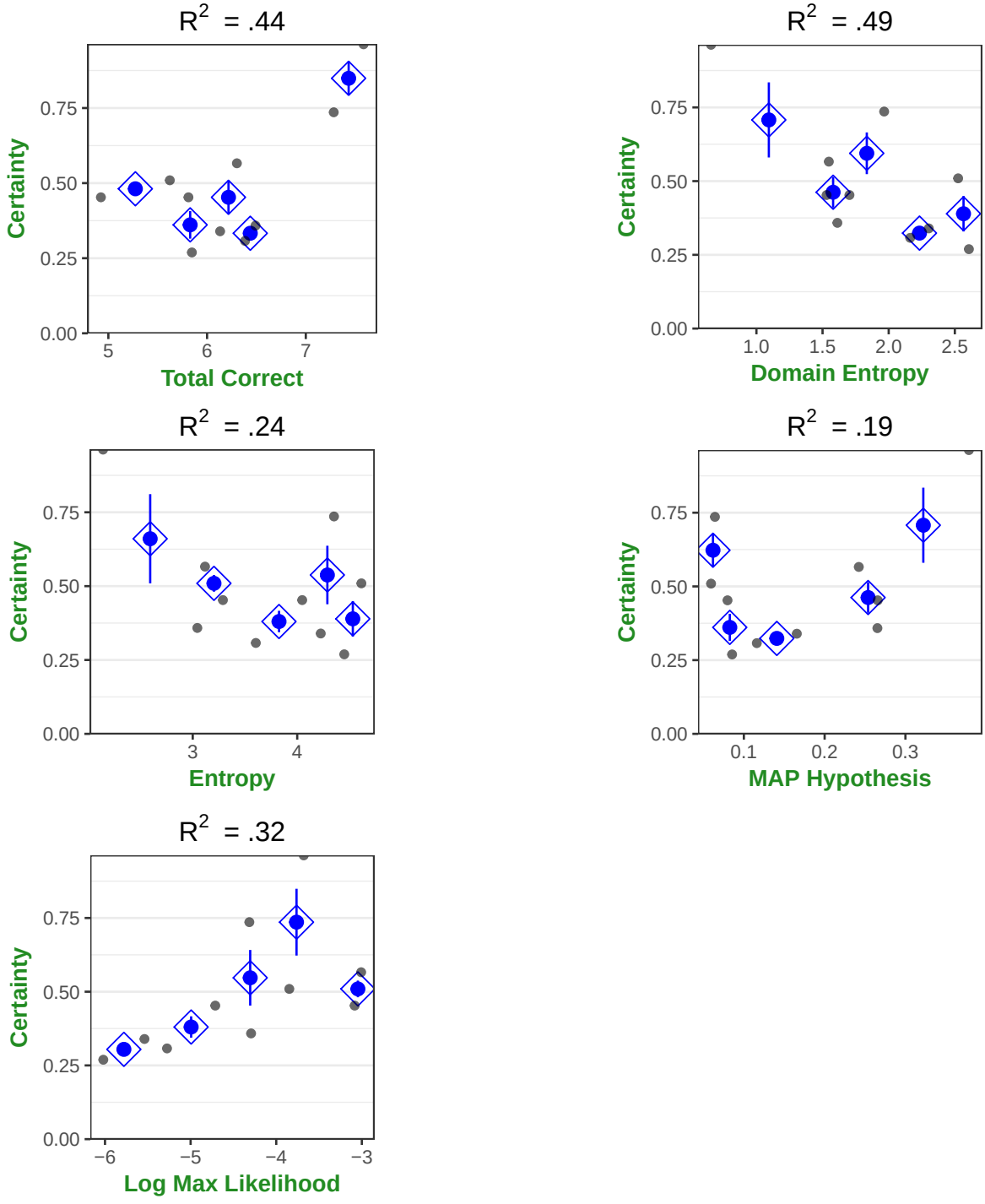


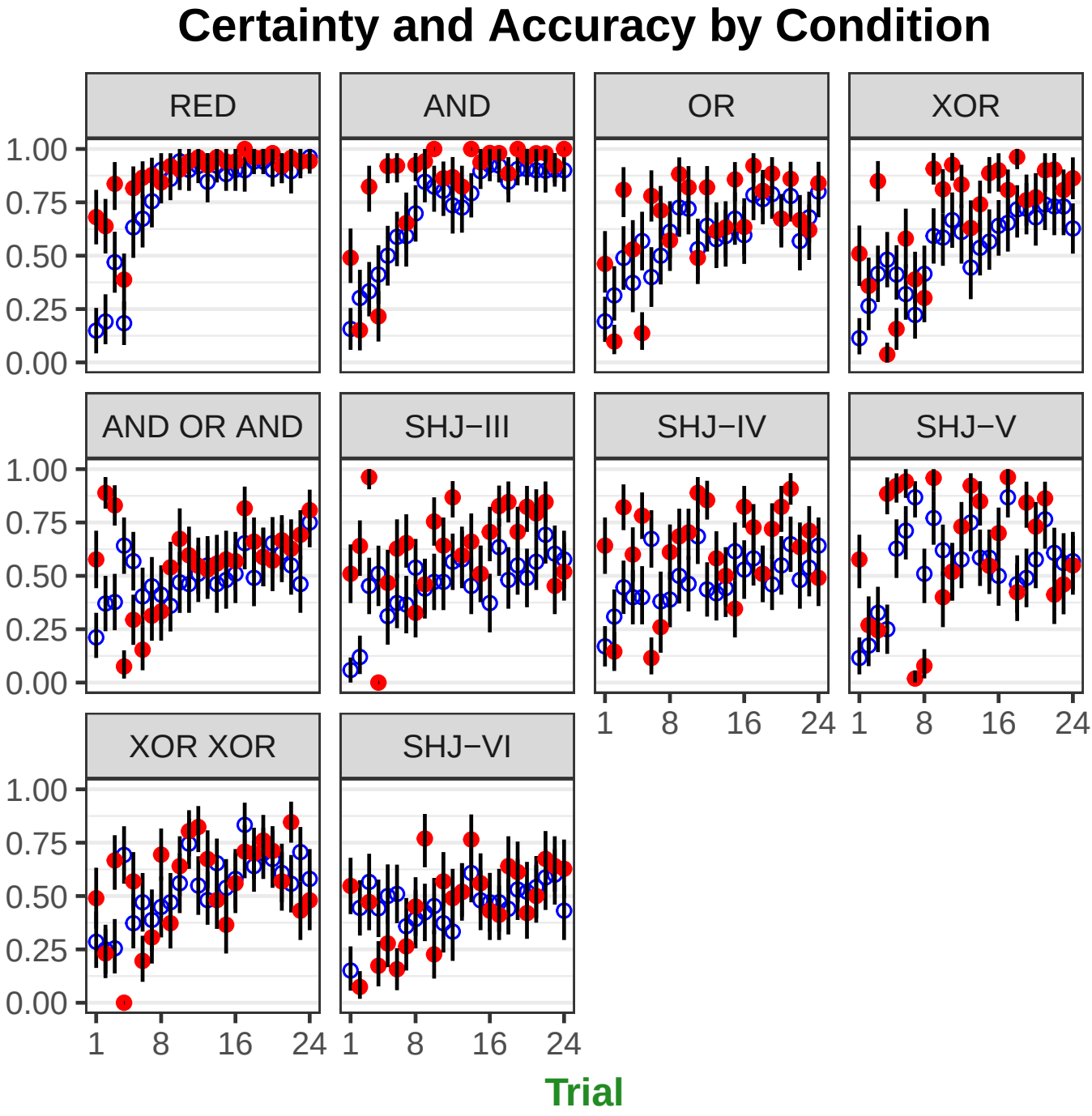
Figure B.3. Key model fits for Experiment 2.

397 **Table C.1.** Predictors of certainty for Experiment 3 (behavioral predictors in gray).

Model	AIC	R^2	Log Likelihood	Beta	Standard Error	p
Local Accuracy 5 Back Current	11617.4	0.70	-5804.7	1.15	0.03	$< .001$
Local Accuracy 5 Back	11700.9	0.68	-5846.5	1.08	0.03	$< .001$
Local Accuracy 4 Back Current	11741.5	0.70	-5866.8	1.10	0.03	$< .001$
Log Local Accuracy 5 Back Current	11748.1	0.65	-5870.0	1.08	0.03	$< .001$
Log Local Accuracy 5 Back	11758.5	0.63	-5875.3	1.03	0.03	$< .001$
Domain Entropy	11767.6	0.61	-5879.8	-1.30	0.04	$< .001$
Log Total Correct	11778.8	0.54	-5885.4	1.00	0.03	$< .001$
Local Accuracy 4 Back	11834.1	0.68	-5913.1	1.02	0.03	$< .001$
Log Local Accuracy 4 Back Current	11878.7	0.65	-5935.4	1.03	0.03	$< .001$
Log Local Accuracy 4 Back	11889.3	0.63	-5940.6	0.98	0.03	$< .001$
Local Accuracy 3 Back Current	11896.4	0.69	-5944.2	1.03	0.03	$< .001$
Total Correct	11909.7	0.50	-5950.9	1.00	0.03	$< .001$
Log Trial	11944.5	0.43	-5968.3	0.89	0.03	$< .001$
Log Max Likelihood	11947.9	0.39	-5970.0	1.16	0.04	$< .001$
Local Accuracy 3 Back	12007.9	0.67	-5999.9	0.94	0.03	$< .001$
Entropy	12022.8	0.43	-6007.4	-1.21	0.04	$< .001$
Log Local Accuracy 3 Back Current	12032.1	0.65	-6012.1	0.97	0.03	$< .001$
Log Local Accuracy 3 Back	12066.6	0.62	-6029.3	0.90	0.03	$< .001$
Trial	12239.6	0.34	-6115.8	0.77	0.03	$< .001$
Local Accuracy 2 Back Current	12243.7	0.64	-6117.8	0.86	0.03	$< .001$
Log Local Accuracy 2 Back Current	12355.0	0.60	-6173.5	0.80	0.03	$< .001$
Local Accuracy 2 Back	12358.1	0.61	-6175.0	0.77	0.03	$< .001$
Log Local Accuracy 2 Back	12404.0	0.57	-6198.0	0.74	0.03	$< .001$
Local Accuracy 1 Back Current	12657.5	0.53	-6324.8	0.63	0.03	$< .001$
MAP	12685.5	0.24	-6338.7	0.80	0.03	$< .001$
Log Local Accuracy 1 Back Current	12729.3	0.48	-6360.6	0.59	0.03	$< .001$
Log MAP	12734.5	0.19	-6363.2	0.69	0.03	$< .001$

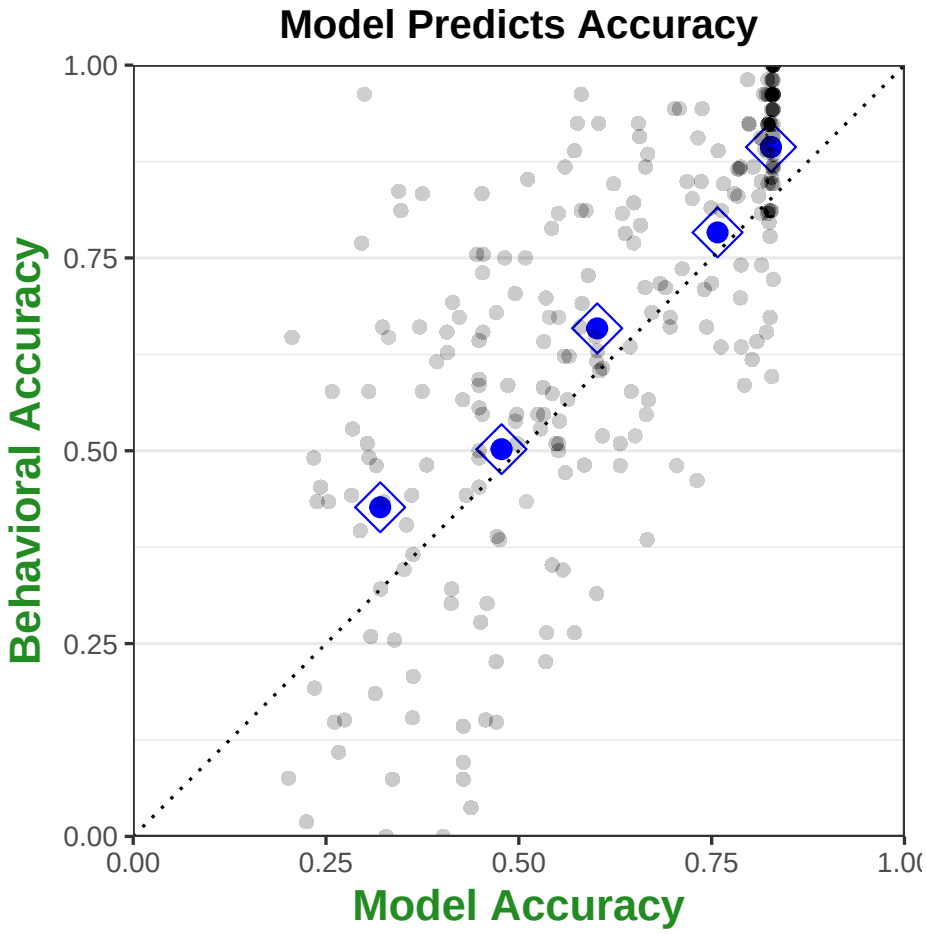
398 **Table C.2.** Predictors of certainty for Experiment 3 using simplified grammar (behavioral predictors in gray).

Model	AIC	R^2	Log Likelihood	Beta	Standard Error	p
Local Accuracy 5 Back Current	11617.4	0.70	-5804.7	1.15	0.03	$< .001$
Local Accuracy 5 Back	11700.9	0.68	-5846.5	1.08	0.03	$< .001$
Local Accuracy 4 Back Current	11741.5	0.70	-5866.8	1.10	0.03	$< .001$
Log Local Accuracy 5 Back Current	11748.1	0.65	-5870.0	1.08	0.03	$< .001$
Log Local Accuracy 5 Back	11758.5	0.63	-5875.3	1.03	0.03	$< .001$
Log Total Correct	11778.8	0.54	-5885.4	1.00	0.03	$< .001$
Local Accuracy 4 Back	11834.1	0.68	-5913.1	1.02	0.03	$< .001$
Log Local Accuracy 4 Back Current	11878.7	0.65	-5935.4	1.03	0.03	$< .001$
Log Local Accuracy 4 Back	11889.3	0.63	-5940.6	0.98	0.03	$< .001$
Local Accuracy 3 Back Current	11896.4	0.69	-5944.2	1.03	0.03	$< .001$
Total Correct	11909.7	0.50	-5950.9	1.00	0.03	$< .001$
Log Trial	11944.5	0.43	-5968.3	0.89	0.03	$< .001$
Local Accuracy 3 Back	12007.9	0.67	-5999.9	0.94	0.03	$< .001$
Log Local Accuracy 3 Back Current	12032.1	0.65	-6012.1	0.97	0.03	$< .001$
Log Local Accuracy 3 Back	12066.6	0.62	-6029.3	0.90	0.03	$< .001$
Log Max Likelihood	12110.3	0.34	-6051.1	0.85	0.03	$< .001$
Trial	12239.6	0.34	-6115.8	0.77	0.03	$< .001$
Local Accuracy 2 Back Current	12243.7	0.64	-6117.8	0.86	0.03	$< .001$
Log Local Accuracy 2 Back Current	12355.0	0.60	-6173.5	0.80	0.03	$< .001$
Local Accuracy 2 Back	12358.1	0.61	-6175.0	0.77	0.03	$< .001$
Log Local Accuracy 2 Back	12404.0	0.57	-6198.0	0.74	0.03	$< .001$
Domain Entropy	12458.0	0.37	-6225.0	-0.84	0.03	$< .001$
Local Accuracy 1 Back Current	12657.5	0.53	-6324.8	0.63	0.03	$< .001$
Log Local Accuracy 1 Back Current	12729.3	0.48	-6360.6	0.59	0.03	$< .001$
Local Accuracy 1 Back	12817.4	0.45	-6404.7	0.50	0.02	$< .001$
Log Local Accuracy 1 Back	12817.4	0.43	-6404.7	0.50	0.02	$< .001$
Max Likelihood	12925.2	0.06	-6458.6	0.48	0.03	$< .001$



399 **Figure C.1.** Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 3. Chance is 50% across all conditions if guesses are

400 made randomly.



401 **Figure C.2.** Model vs. behavioral accuracy for Experiment 3.

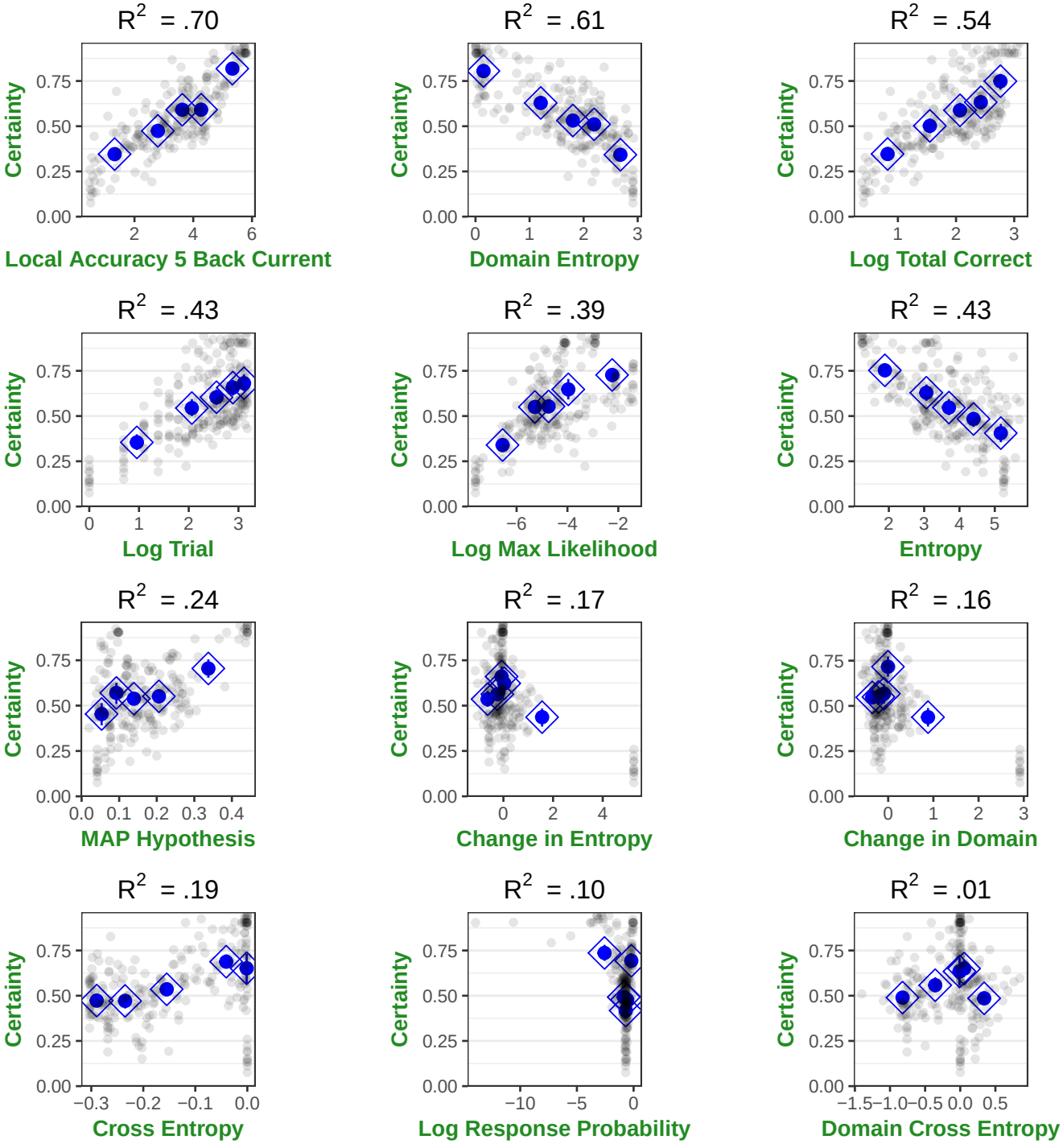


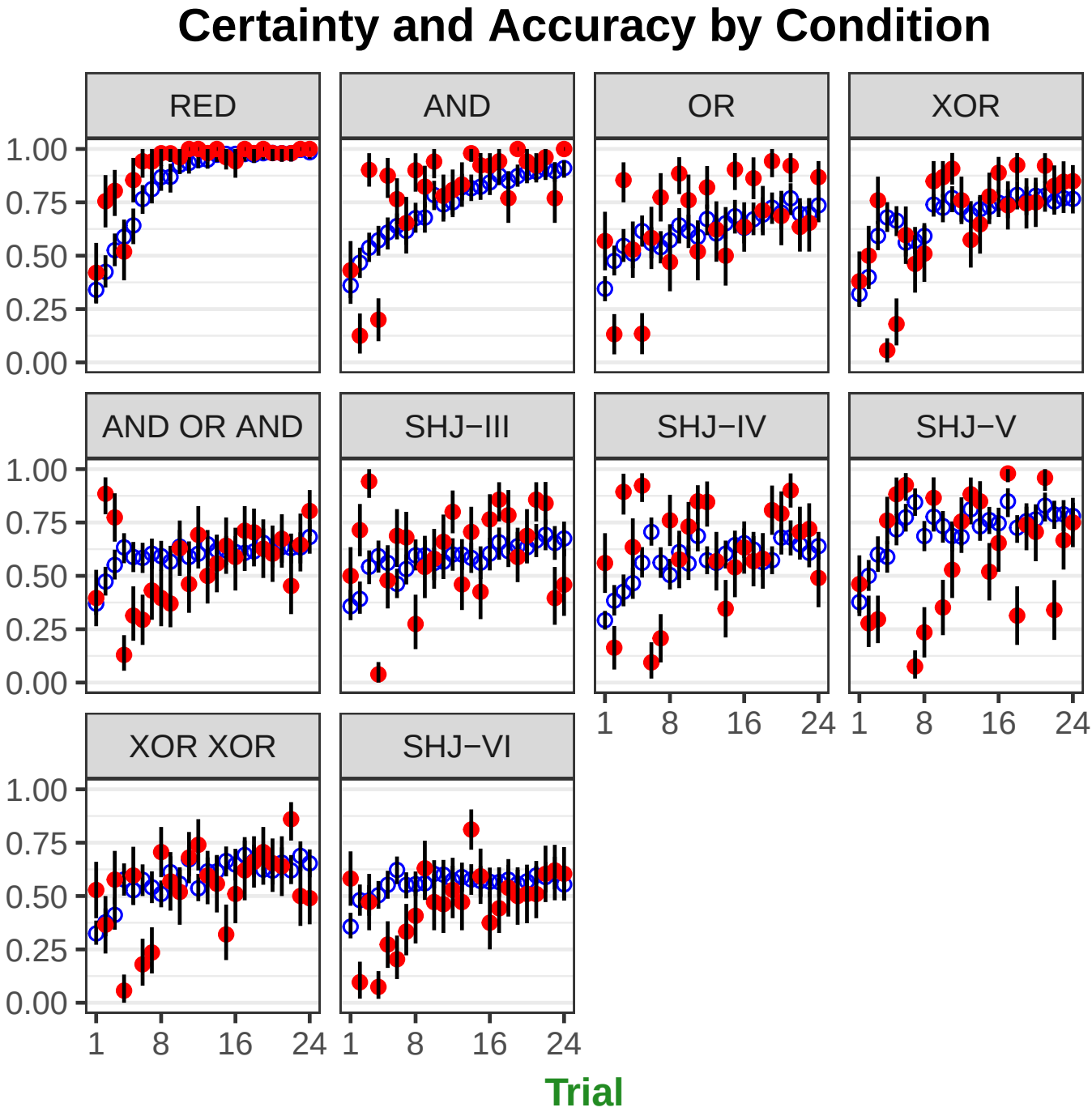
Figure C.3. Key model fits for Experiment 3.

403 **Table D.1.** Predictors of certainty for Experiment 4 (behavioral predictors in gray).

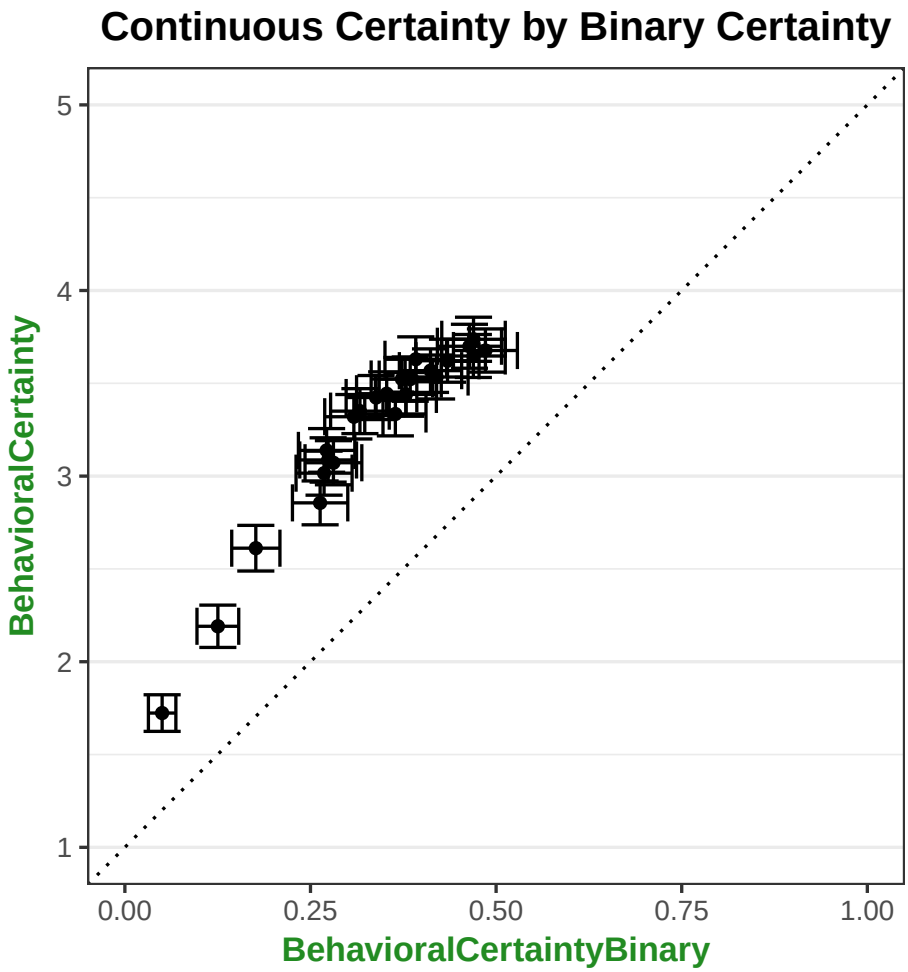
Model	AIC	R^2	Log Likelihood	Beta	Standard Error	p
Local Accuracy 5 Back Current	33739.2	0.77	-16864.6	0.59	0.01	$< .001$
Local Accuracy 5 Back	33749.9	0.76	-16870.0	0.57	0.01	$< .001$
Log Total Correct	33912.8	0.56	-16951.4	0.52	0.01	$< .001$
Log Local Accuracy 5 Back	34008.6	0.67	-16999.3	0.52	0.01	$< .001$
Local Accuracy 4 Back Current	34096.7	0.77	-17043.3	0.56	0.01	$< .001$
Domain Entropy	34102.2	0.69	-17046.1	-0.63	0.01	$< .001$
Local Accuracy 4 Back	34108.8	0.76	-17049.4	0.54	0.01	$< .001$
Log Local Accuracy 5 Back Current	34121.6	0.70	-17055.8	0.53	0.01	$< .001$
Log Trial	34130.0	0.44	-17060.0	0.47	0.01	$< .001$
Total Correct	34231.1	0.54	-17110.5	0.51	0.01	$< .001$
Log Local Accuracy 4 Back	34317.0	0.68	-17153.5	0.50	0.01	$< .001$
Log Max Likelihood	34388.4	0.41	-17189.2	0.58	0.01	$< .001$
Log Local Accuracy 4 Back Current	34453.8	0.70	-17221.9	0.51	0.01	$< .001$
Entropy	34570.0	0.52	-17280.0	-0.62	0.01	$< .001$
Local Accuracy 3 Back Current	34574.8	0.76	-17282.4	0.52	0.01	$< .001$
Local Accuracy 3 Back	34601.6	0.74	-17295.8	0.49	0.01	$< .001$
Log Local Accuracy 3 Back	34787.0	0.68	-17388.5	0.47	0.01	$< .001$
Log Local Accuracy 3 Back Current	34910.0	0.70	-17450.0	0.47	0.01	$< .001$
Trial	34919.9	0.33	-17455.0	0.41	0.01	$< .001$
Local Accuracy 2 Back Current	35311.1	0.70	-17650.6	0.44	0.01	$< .001$
Local Accuracy 2 Back	35348.4	0.67	-17669.2	0.41	0.01	$< .001$
Log Local Accuracy 2 Back	35469.7	0.62	-17729.9	0.39	0.01	$< .001$
Log Local Accuracy 2 Back Current	35554.7	0.65	-17772.4	0.40	0.01	$< .001$
MAP	35855.3	0.33	-17922.6	0.46	0.01	$< .001$
Log MAP	35939.9	0.26	-17964.9	0.41	0.01	$< .001$
Local Accuracy 1 Back Current	36165.2	0.56	-18077.6	0.31	0.01	$< .001$
Change in Entropy	36191.8	0.16	-18090.9	-0.28	0.01	$< .001$

404 **Table D.2.** Predictors of certainty for Experiment 4 using simplified grammar (behavioral predictors in gray).

Model	AIC	R^2	Log Likelihood	Beta	Standard Error	p
Local Accuracy 5 Back Current	33739.2	0.77	-16864.6	0.59	0.01	$< .001$
Local Accuracy 5 Back	33749.9	0.76	-16870.0	0.57	0.01	$< .001$
Log Total Correct	33912.8	0.56	-16951.4	0.52	0.01	$< .001$
Log Local Accuracy 5 Back	34008.6	0.67	-16999.3	0.52	0.01	$< .001$
Local Accuracy 4 Back Current	34096.7	0.77	-17043.3	0.56	0.01	$< .001$
Local Accuracy 4 Back	34108.8	0.76	-17049.4	0.54	0.01	$< .001$
Log Local Accuracy 5 Back Current	34121.6	0.70	-17055.8	0.53	0.01	$< .001$
Log Trial	34130.0	0.44	-17060.0	0.47	0.01	$< .001$
Total Correct	34231.1	0.54	-17110.5	0.51	0.01	$< .001$
Log Local Accuracy 4 Back	34317.0	0.68	-17153.5	0.50	0.01	$< .001$
Log Local Accuracy 4 Back Current	34453.8	0.70	-17221.9	0.51	0.01	$< .001$
Local Accuracy 3 Back Current	34574.8	0.76	-17282.4	0.52	0.01	$< .001$
Log Max Likelihood	34585.0	0.35	-17287.5	0.46	0.01	$< .001$
Local Accuracy 3 Back	34601.6	0.74	-17295.8	0.49	0.01	$< .001$
Log Local Accuracy 3 Back	34787.0	0.68	-17388.5	0.47	0.01	$< .001$
Log Local Accuracy 3 Back Current	34910.0	0.70	-17450.0	0.47	0.01	$< .001$
Trial	34919.9	0.33	-17455.0	0.41	0.01	$< .001$
Local Accuracy 2 Back Current	35311.1	0.70	-17650.6	0.44	0.01	$< .001$
Local Accuracy 2 Back	35348.4	0.67	-17669.2	0.41	0.01	$< .001$
Domain Entropy	35357.0	0.47	-17673.5	-0.47	0.01	$< .001$
Log Local Accuracy 2 Back	35469.7	0.62	-17729.9	0.39	0.01	$< .001$
Log Local Accuracy 2 Back Current	35554.7	0.65	-17772.4	0.40	0.01	$< .001$
Local Accuracy 1 Back Current	36165.2	0.56	-18077.6	0.31	0.01	$< .001$
Local Accuracy 1 Back	36267.8	0.49	-18128.9	0.28	0.01	$< .001$
Log Local Accuracy 1 Back	36267.8	0.46	-18128.9	0.28	0.01	$< .001$
Log Local Accuracy 1 Back Current	36295.2	0.51	-18142.6	0.29	0.01	$< .001$
Max Likelihood	36367.0	0.08	-18178.5	0.28	0.01	$< .001$



405 **Figure D.1.** Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 4. Chance is 50% across all conditions if guesses are
406 made randomly.



407 **Figure D.2.** Continuous certainty (Experiment 4) and binary certainty (Experiment 1) grouped by trial.

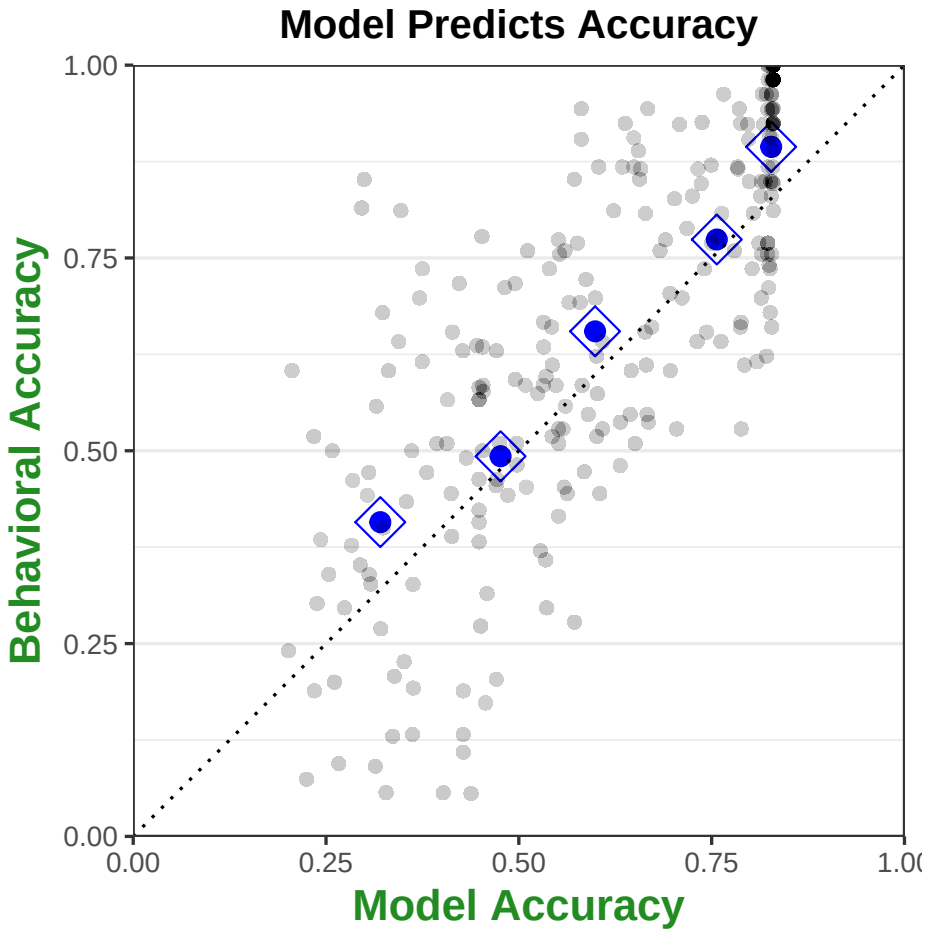
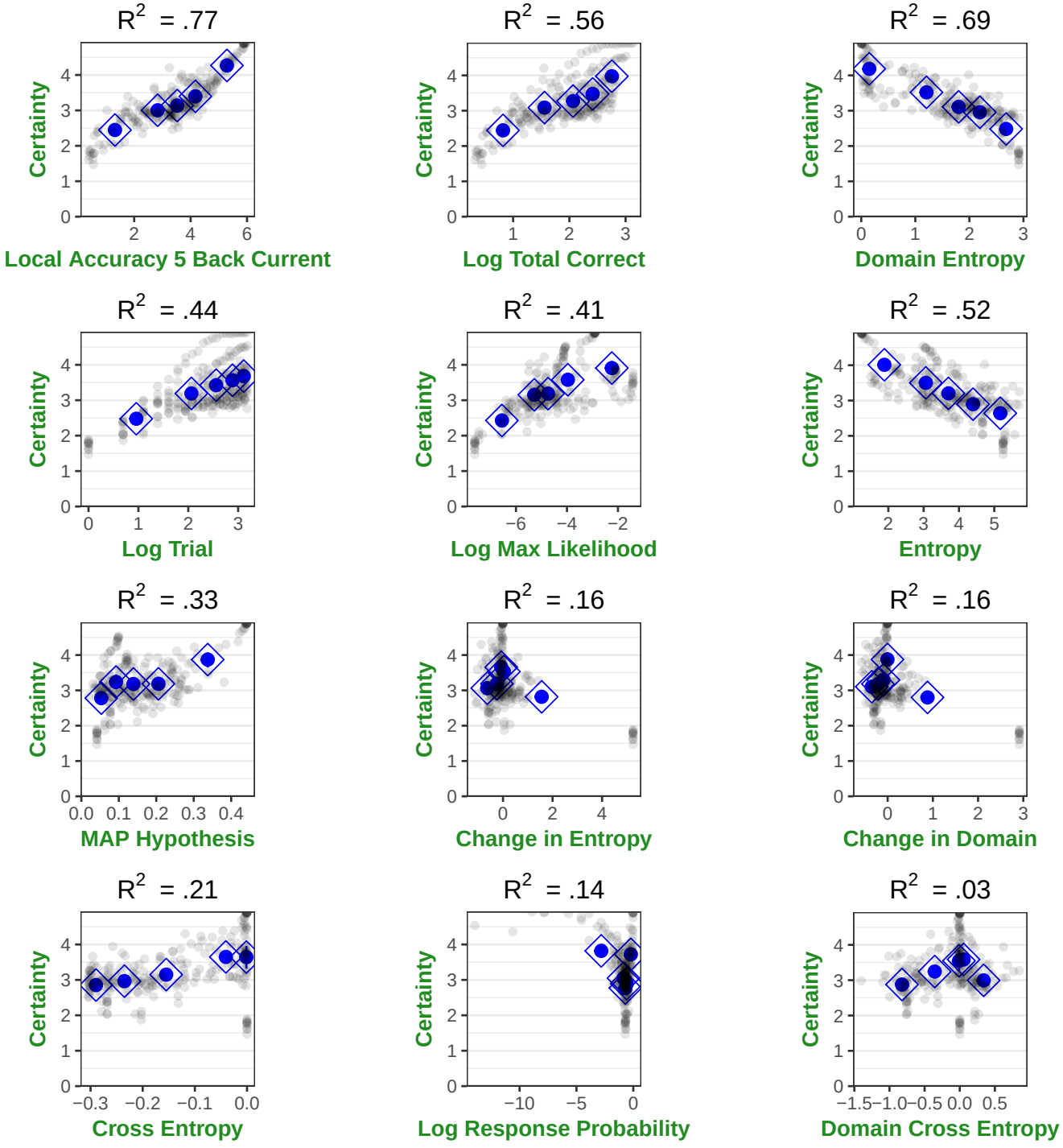


Figure D.3. Model vs. behavioral accuracy for Experiment 4.



409 **Figure D.4.** Key model fits for Experiment 4.