Title
Certainty and the Source of Misinformed Beliefs

Author
Marti, Louis

Publication Date
2021

Peer reviewed|Thesis/dissertation

Certainty and the Source of Misinformed Beliefs

By

Louis E. Martí

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Steven Piantadosi, Chair
Assistant Professor Celeste Kidd
Professor Alison Gopnik

Fall 2021

Abstract

Certainty and the Source of Misinformed Beliefs

by

Louis E. Martí

Doctor of Philosophy in Psychology

University of California, Berkeley

Assistant Professor Steven Piantadosi, Chair

Humans possess a metacognitive sense of certainty which, for better or worse, influences behavior. This sense of certainty is often misleading and can leave us vulnerable to believing false information. In this dissertation, I study how humans form their sense of certainty and the types of false beliefs which we can be at times, highly certain of. This work spans across multiple domains, including concept learning, word-meaning, pseudoscience, and people's metacognitive beliefs. Across seven experiments, I present empirical evidence that learners use heuristics over idealized model-based features when forming their sense of certainty, and that this leaves us prone to errors which can result in the adoption of misinformed beliefs as drastic as the belief that the Earth is flat.

Table of Contents

I. Introduction

How do humans believe untruths?

Intelligent organisms possess abilities which allow them to assess their own thought processes. These abilities are collectively known as metacognition. Metacognition allows a wide variety of abilities such as analyzing the state of someone's knowledge and strategizing over learning methods (Flavell, 1979). One metacognitive signal that guides our learning is our sense of certainty (e.g., Martí et al., 2018; Wade & Kidd, 2019), which provides feedback used to determine the strength of our beliefs and help us update our knowledge states. This sense of certainty can itself be inaccurate which, in some cases, can lead humans to be strongly certain about beliefs which are untrue. These beliefs can range from an innocuous falsehood such as "it will be cloudy tomorrow" to more dangerous beliefs such as "vaccines cause autism" or "positive thinking can cure my cancer".

In this dissertation, I investigate the human sense of certainty using behavioral experimentation and Bayesian models in order to understand how humans believe inaccuracies about the world. Specifically, I investigate what makes people certain, whether those feelings of certainty can be miscalibrated even in a routine domain such as shared word-meanings, and whether specific misinformed beliefs create a "slippery slope" into other misinformed beliefs. I present empirical evidence that our sense of certainty is formulated primarily using imprecise heuristics which, though often "good enough", can also lead humans to stubbornly held inaccurate beliefs.

How do we form our sense of certainty?

Past work examining the sense of certainty in humans have resulted in a competing series of frameworks that have had varying degrees of success across different domains. Signal Detection Theory (Peterson et al., 1954; Tanner & Swets, 1954) models certainty as a representation of the noise around a stimulus. If one views a vase full of marbles, they might estimate it contains 110 marbles. If, for example, this is done dozens of times, the resulting series of estimates will have a standard deviation which is equal to the individual's certainty. Following the introduction of Signal Detection Theory, accumulator models were developed which took into account the amount of time which was available for a given decision (De Martino, et al., 2013). By incorporating time, these models help explain empirical evidence that the amount of time we have to make a decision is correlated with our sense of certainty about that decision.

Models based on Signal Detection Theory treat our sense of certainty as being derived solely from intrinsic features of the stimuli. If an image is blurry, people will be uncertain about what it is; if it is clear, they will be certain. Yet there exists an entire body of research which points to our sense of certainty as incorporating more factors than the intrinsic features of stimuli, and possessing the capacity to incorporate extraneous factors under certain conditions. A review of individuals as young as middle-schoolers taking methylphenidate and amphetamine found that while they believe taking stimulants helped them improve on certain tasks, they performed no better than controls (Smith & Farah, 2011). In one such study, working memory was assessed by presenting participants with a series of numbers or letters, followed by a series of probe items where participants responded whether or not they were in the original set. Participants given 5, 10, or 20 mg of methylphenidate performed no better than placebo controls (Callaway, 1983). Even superficial task features such as characteristics of a printed typeface can influence certainty about high-level decisions such as whether an individual is guilty of a crime. Participants were presented with either positive or negative witness testimony followed by objective case facts in either a fluent or disfluent font. Positive or negative bias, and the certainty about that bias, was either maintained (with fluent font) or overridden (with disfluent font) (Hernandez & Preston, 2013). Additionally, simply presenting more information will increase certainty, irrespective of any effects on task performance. When participants were asked to guess the outcomes of specific football games, their accuracy did not improve when given more information about past game outcomes, but their certainty increased (Tsai et al., 2008).

These findings have informed alternatives to Signal Detection Theory which examine the effects of a host of heuristics and cognitive biases on certainty. These biases are argued to have arisen as a way for our finite mental resources to deal with the overwhelming amount of information our cognition has to deal with. Two such examples are the representative heuristic and the availability heuristics (Tversky & Kahneman, 1974). The representative heuristic occurs when we compare a situation to an existing mental prototype. If everyone I know who is a Virgo is shy, I might think a stranger I meet who is a Virgo is also shy, even though this correlation is entirely coincidental. Similarly, the availability heuristic refers to our tendency to overuse or overvalue information that immediately comes to mind. My belief that there are regular homicides in my community will strengthen if the local news frequently reports on a single but particularly sensational murder.

The reliance on heuristics over more objective evidence has been the focus of dual process theory which describes two distinct thought processes that humans engage in. Type 1 processes are error-prone but also fast, automatic, and intuitive, and are typically considered to be our default mode while going about our daily activities. Type 2 processes on the other hand are slow, cognitively demanding, and reflective

(Wason & Evans, 1974). Reasoning errors have been identified both due to a failure to engage in Type 2 reasoning (Evans, 2007), and a failure of Type 2's results to override Type 1's results (De Neys, 2012). There is also evidence that both forms of thinking occur in parallel (Sloman, 1996) but Type 1 is what is predominantly acted upon because it is substantially faster than Type 2 (Handley & Trippas, 2015). Research into individual differences has found differing time ratios of thinking type (Stanovich & West, 2000). These differences are not only due to an individual's ability to engage in Type 2 thinking, but also in their willingness to (Stanovich, 2004). For example, people who are more intellectually humble, and realize they might be incorrect, engage in more Type 2 thinking (Baron, 2008). Yet despite these individual differences, Type 2 thinking can take center stage due to simple interventions such as task instructions (Daniel & Klaczynski, 2006) and increasing response time (Evans & Curtis-Holmes, 2005). Interestingly, the mental-health literature has consistently found a positive correlation between cognitive biases, false beliefs, and *positive* mental health outcomes (Lefcourt, 1973; Taylor, 1989). This indicates that Type 1 thinking is adaptive, not only due to speed and resource conservation, but because of mental health as well.

What leads to misinformed beliefs?

Misinformed beliefs range from "I bought milk at the grocery store yesterday." to "water memory will cure my cancer" or "the Earth is flat". Fake news can lead to the implantation of false memories from simply reading a fake news article (Murphy, et al., 2019). Individuals also reliability rate themselves as less susceptible to cognitive biases compared to the "average American". This effect persisted even when they were presented with evidence that cognitive biases are pervasive and often go undetected by the individual (Pronin et al., 2002). People also believe that their skills can help them succeed in tasks that are purely governed by chance, like predicting the results of tossing a fair coin (Langer & Roth, 1975), and that an entire host of ineffective medical treatments can cure their ailments (Matute et al., 2011). These types of beliefs have been argued to be self-serving, and even self-protecting (Taylor & Brown, 1988). The illusion of control, the idea that one has more control over an outcome than one actually does (Langer, 1975), has been negatively associated with depression. Specifically, depressed individuals tend to show little to no illusions of control, while non-depressed individuals happily live with their illusions (Abramson et al., 1978).

Another finding likely related to the availability heuristic has been named the illusory truth effect. This effect occurs when the mere repetition of information (false or not) increases human certainty about the truth of that information. In one study participants reported certainty ratings about the truthfulness of statements regarding politics, sports, and the arts. This was done over three sessions with two week intervals between them. Statements which were repeated between sessions were found to increase in certainty over time (Hasher et al., 1977). The same effect occurs even if the

presented information is explicitly labeled as an opinion (Arkes, et al., 1989). This finding is so ubiquitous that it occurs even when controlling for cognitive ability, cognitive closure (aversion to ambiguity), and cognitive style (De keersmaecker et al., 2019). There is also evidence that this bias may be present at birth given that it has already been found in five-year-olds to be just as strong as in adults. In a single session, five-year-olds, ten-year-olds, and adults, were presented with both true and untrue nature statements. They were then presented with new statements and some statements which they had seen before and asked to rate their truth. Across all age groups, statements which had been seen before, regardless of their truth, were more likely to be rated as true (Fazio & Sherry, 2020). More recently, there has been evidence that a small minority of individuals possess a negative truth effect where repetition causes a *decrease* in belief strength. Eight separate datasets were examined on an individual-level (as opposed to traditional group-level analyses) and roughly 1% of participants displayed a negative truth effect (Schnuerch et al., 2020).

Inaccurate beliefs have been found to be associated with a host of personality traits or thinking errors. Conspiracy beliefs are modestly associated with agreeableness and conscientiousness as personality traits (Bowes, et al., 2021), a conservative political ideology, and paranoid ideation (van der Linden, et al., 2021). Lastly, self-described open-minded thinkers tend to believe less pseudoscience such as climate-change denial, extrasensory perception, and the paranormal (Pennycook, et al., 2020). Individuals who become highly certain with very little evidence are predisposed to pseudoscientific beliefs. Sanchez and Dunning presented participants with a game where the goal was to guess which out of two lakes was being fished from. One lake had predominantly gray fish while the other primarily had orange fish. A fisherman would catch a fish, show the color, and would continue catching or stop depending on the participant's wishes. Some individuals only needed one or two catches before they became certain they knew what lake it was, despite insufficient evidence (Sanchez & Dunning, 2020). Pseudoscientific beliefs are also associated with the illusion of causality, or the mistaken belief that a causal connection exists when it does not (Torres et al., 2020).

Bayesian modeling

More recently, Bayesian approaches to cognition have introduced the idea that our beliefs (including certainty) are a result of accumulated prior knowledge combined with a likelihood calculation of hypotheses given the observed data. Probability-based models provide a useful framework to formalize uncertainty. In a Bayesian learner model, the probability of a given hypothesis represents the belief strength of that hypothesis. Certainty can then be explicitly formalized as either the distance from a probability of one, or as the entropy over the entire hypothesis space (Glymour, 2003).

These types of frameworks have been used to model word learning in children (Xu & Tenenbaum, 2007), along with curiosity and exploration (Gopnik & Bonawitz, 2015), among many other domains.

The primary research paradigm I employ is the combination of behavioral experimentation and computational modeling. This combination allows me to rigorously test theories and predictions against empirical data. For example, in testing whether human certainty is calibrated to reality in Chapters 2 and 5, a Bayesian ideal learner model allows me to simulate the task and calculate multiple plausible predictors of certainty. These different types of certainty, in conjunction with purely behaviorally calculated certainties, can be compared to the self-reported certainties of participants in order to discover which types of certainty best predict human behavior. The experiments in Chapter 3 use a non-parametric Bayesian clustering model which receives behavioral data as input, and outputs a distribution of data clusters. By applying this to individual differences in word-meaning, the likely number of unique concepts present in our sample can be recovered. These results are then used as input into an ecological species estimator which allows me to estimate the total number of unique concepts on a population level. Chapter 4 uses demographic data and a process known as iterative proportional fitting to estimate the true prevalence of misinformed beliefs in the U.S.

The predictors of adult certainty

Past research has discovered that while certainty seems to be well calibrated in low-level perceptual domains, it is not well calibrated in high-level domains such as concept learning or in complex belief networks such as misinformed beliefs. For example, subjective uncertainty regarding visual stimuli reliably predicts objective uncertainty. Participants viewed either pairs of Gabor patches or pairs of alphabetic symbols with visual noise and reliably chose the option with less uncertainty. (Barthelmé & Mamassian, 2009). Similarly, objective uncertainty reliably predicts subjective uncertainty for auditory stimuli and in numerical discrimination tasks (Sanders et al., 2016). These findings are in contrast to high-level domains such as social interactions, where simply hearing that opinions are shared by others raises certainty, even if those opinions are inaccurate (Yaniv et al., 2009). Likewise, if an expert contradicts preexisting beliefs, individuals tend to increase their certainty in that belief (Tormala et al., 2011; Tormala & Petty, 2004).

Understanding how certainty is often inaccurate cannot be done without first knowing how our sense of certainty is calculated. One step towards understanding this calculation is discovering predictors which accurately predict people's sense of certainty. Chapter 2 outlines a series of experiments which, for the first time, test the prediction strength of both model-based and behavioral predictors of certainty where participants were asked to self-report their certainty while learning a high-level Boolean concept. Crucially the predictors were all part of one of two classes of predictors. Model-based

predictors were derived from a Bayesian ideal learner model and were based on criteria such as the entropy over learned concepts. Behavioral predictors on the other hand were based on superficial task features such as the number of trials a participant guessed correctly. We found that while both types of predictors uniquely predict certainty, a behavioral predictor, the number of correct trials in the recent past, predicted certainty the best. These results suggest that in high-level domains such as concept learning, humans tend to use “good enough” heuristics over veridical task features to inform their certainty.

Shared concepts across humans

While the certainty study provided evidence as to how humans can arrive at false beliefs, my next goal was to examine false beliefs in a more real-world domain than abstract Boolean concepts. To facilitate this, I asked whether real-world concepts are shared between humans for a given word-sense, and whether people’s certainty is correctly calibrated to the amount they are shared.

Past research into individual differences in word-meaning has found some diversity, starting with variance in the way individuals classify cups and bowls depending on height and width (Labov, 1973), which was later shown to extend to natural categories in general (McCloskey & Glucksberg, 1978). Other research has cataloged individual differences in typicality judgements using various natural and artificial categories (Barsalou, 1987; Verheyen & Storms, 2013; Koriat & Sorka, 2015). Yet none have quantified the number of distinct concepts in the population for a given word sense, nor examined people’s certainty about the amount of diversity. Because of this, I ran two experiments where participants were asked to give similarity judgements or feature ratings about either common animals or well-known politicians. Participants were also asked to guess the number of other individuals who would give the same response. Using a Bayesian clustering model, and a species ecological estimator, I estimated the existence of roughly a dozen concepts per word-sense. There was also a significant miscalibration in people’s perception of the true amount of diversity present in the sample. Specifically, individuals tend to display a strong egocentric bias, overestimating the number of people who share their concept. These results suggest that the findings described in Chapter 2 likely extend to everyday concepts given that participants substantially overestimate how shared their concepts are.

The prevalence of misinformed beliefs

Chapters 2 and 3 provide evidence that simple Boolean concepts and even relatively complex everyday concepts are subject to a miscalibration of certainty due to a reliance on heuristics. Chapter 4 seeks to extend ecological validity even further by examining misinformed complex high-level belief systems such as COVID-19 conspiracy theories. In particular, my goal was to test the hypothesis that in contrast to

being solely in the realm of fringe conspiracy theorists, misinformed beliefs are broadly spread throughout the population. Research on misinformed beliefs has traditionally focused on characterizing the population of a single belief. For example, catalogs have been made of flat Earthers (Landrum, 2021), anti-vaxxers (Martinez-Berman et al., 2021), and climate change deniers (Uscinski et al., 2017). However, due to their singular focus, these studies resulted in the appearance of a subpopulation of seemingly susceptible individuals. In order to assess whether this was truly the case, I collected participant certainty ratings about 30 different misinformed beliefs. In contrast to there being a relatively small population of particularly gullible individuals, the data showed that these misinformed beliefs are broadly, but thinly, spread across the entire population with the median participant believing in 9 out of 30 misinformed beliefs. There was also modest evidence that certain beliefs were associated with other beliefs, which implies the existence of a slippery slope in some cases and points to a higher-order belief structure. Together, Chapters 2, 3, and 4 provide evidence that everyone's sense of certainty can quite easily lead them away from truth, possibly due to an over reliance on heuristics.

II. Certainty Is Primarily Determined by Past Performance During Concept Learning

Louis Martí, Francis Mollica, Steven Piantadosi, & Celeste Kidd

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) and Tormala and Petty (2004) found that when people were 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. Similarly, 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 et al., 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 are equally as confident regardless of accuracy. This applies to both answers to questions and predictions about the accuracy of others (Prelec et al., 2017). Additionally, confidence in a memory has no relationship to whether or not the memory actually occurred (Loftus et al., 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 et al., 2014; Kepecs, et al., 2008; Kiani et al., 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 while focusing on a manageable, simplified space of hypotheses (Bruner & Austin, 1986; Feldman, 2000; Goodman et al., 2008; Shepard et al., 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

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 2-1 and Table 2-1). 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 Appendix II-A for additional method details).

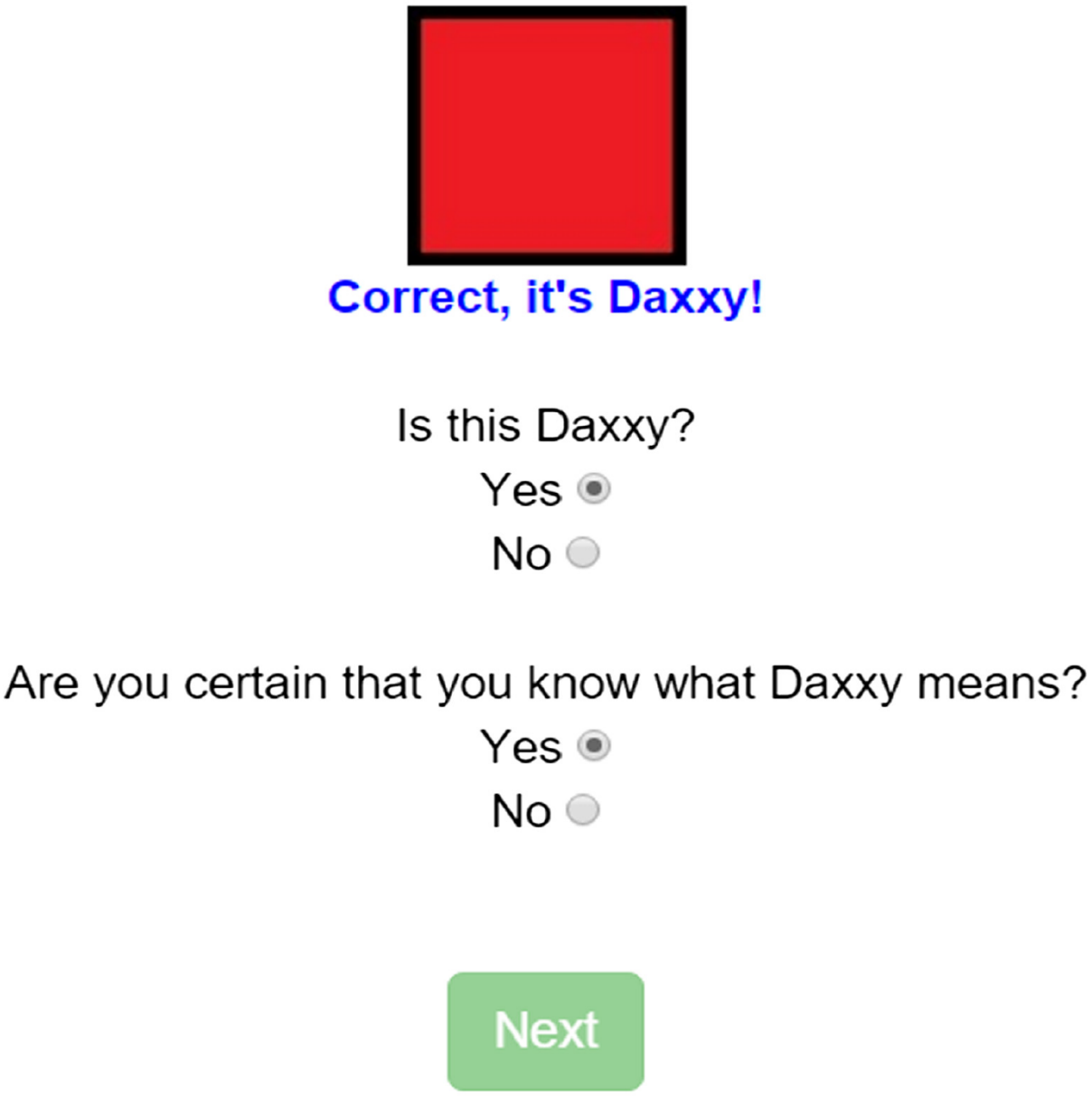

Figure 2-1: In Experiment 1, participants saw 24 trials (as above), randomized between conditions. Feedback was displayed after responding.

		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

Table 2-1: Concepts presented to participants. Concepts 1 and 5–9 are the Shepard, Hovland, and Jenkins family consisting of three features and four positive examples.

Results

We first visualize plots of participants' certainty and accuracy for each concept in order to show (a) whether certainty and accuracy improved over the course of the experiment, (b) whether theoretically harder concepts (according to Feldman, 2000) were, in fact, more difficult for participants, and (c) whether participants' certainty correlated with their accuracy in general.

Figure 2-2 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.

Predictor	Description
Trial	Number of trials seen so far
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	Model uncertainty over hypotheses regarding what the concept is
Domain Entropy	Model 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	The probability of the participant's response given the model predictions

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

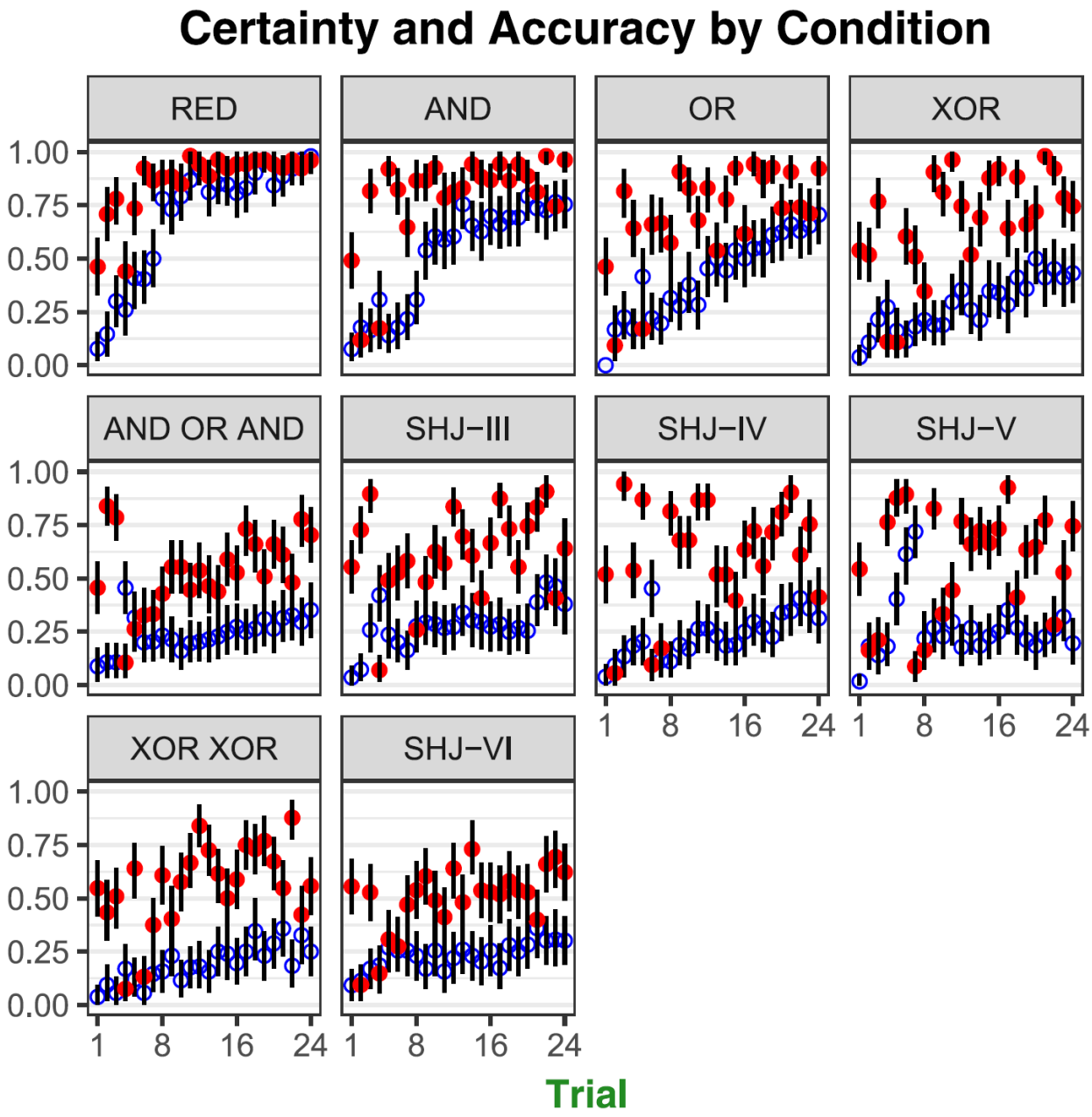

Figure 2-2: Mean certainty (hollow blue circles) and mean accuracy (filled red circles) across concepts for Experiment 1. Chance is 50% across all conditions if guesses are made randomly.

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

[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 2-4 in the Appendices.

shows model accuracy significantly predicts behavioral accuracy (R^2 = .50, β = .748, z = 30.423, p < .001; Figure 2-3), meaning that overall performance can be reasonably well predicted by the learning model.

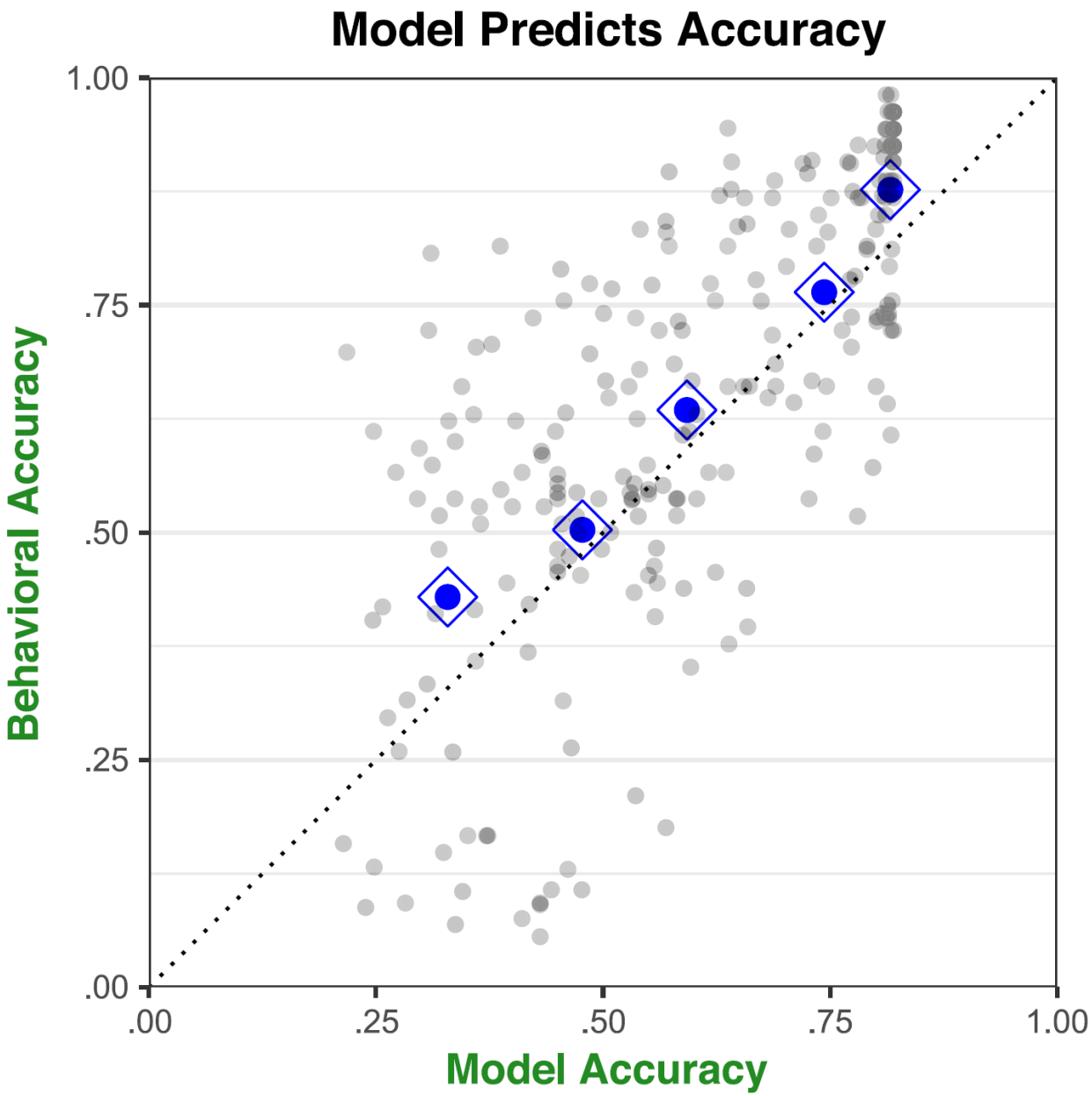

Figure 2-3: Model vs. behavioral accuracy for Experiment 1.

Figure 2-4 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 a high R^2 and very little residual variance. Local Accuracy 5 Back, the accuracy averaged over the past 5 items, has a

high R^2, meaning that individuals with low local accuracy were uncertain and individuals with high local accuracy were highly certain. Likewise, Domain Entropy also has a high $R^2$and is very ordered compared to the other model predictors (see Figure 2-7 in the Appendices for additional predictor visualizations).

Table 2-5 in the Appendices shows the full model results, giving the performance of each model in predicting certainty ratings.[2] These have been sorted by Akaike information criterion (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^2measure, 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 Accuracy models outperform most of the other alternatives, a pattern that 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.

[2] See Table 2-6 in the Appendices for simplified grammar predictors.

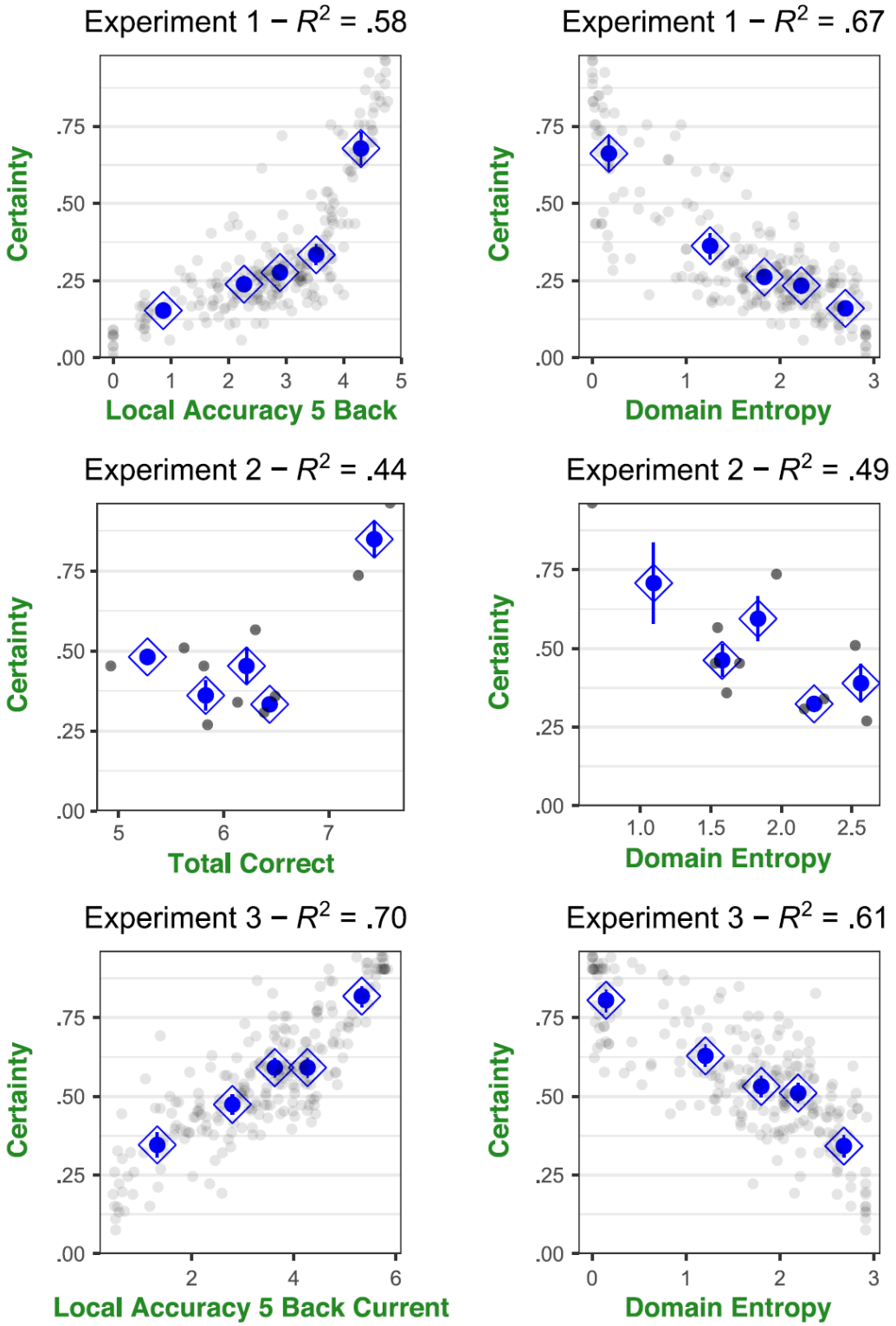

Figure 2-4: Key model fits for Experiments 1-3, showing mean participant responses for each concept and trial (gray) and binned model means in each of five quantiles (blue) for certainty rating (y-axis) as a function of model (x-axis). Diagonal lines with low variance correspond to models which accurately capture human behavior.

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	.054
Log Maximum Likelihood	−0.04	0.04	−1.11	.269

Table 2-3: Regression for best predictors (standardized) in Experiment 1 (behavioral predictors in gray).

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. Strikingly, the idealized model of entropy over hypotheses–what might have corresponded to our best a priori guess for what certainty should reflect—performs especially poorly, worse than many behavioral and other model-based predictors. 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.

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 2-3).[3][4] As the table makes clear, all behavioral predictors, along with Domain Entropy, make significant, unique contributions to certainty. 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 Appendix II-D). A plausible hypothesis would then be that the predictors derived from our ideal learning model would also be related

[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 (Barr, Levy, Scheepers, & Tily, 2013) in this regression due to a lack of convergence.

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^2value, however it is important to emphasize that the R^2values did not take into account the subject and condition used in the mixed effect model. When these effects are controlled, we find that Domain Entropy has less of an influence than behavioral predictors, although its contribution to certainty is still nonzero. Performance of the predictors in a model that controls these effects should be a more reliable guide to each predictor's effect. 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 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 2-5 and Figure 2-6). 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 2-5). 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 2-6). (See Appendix II-B for further details.)

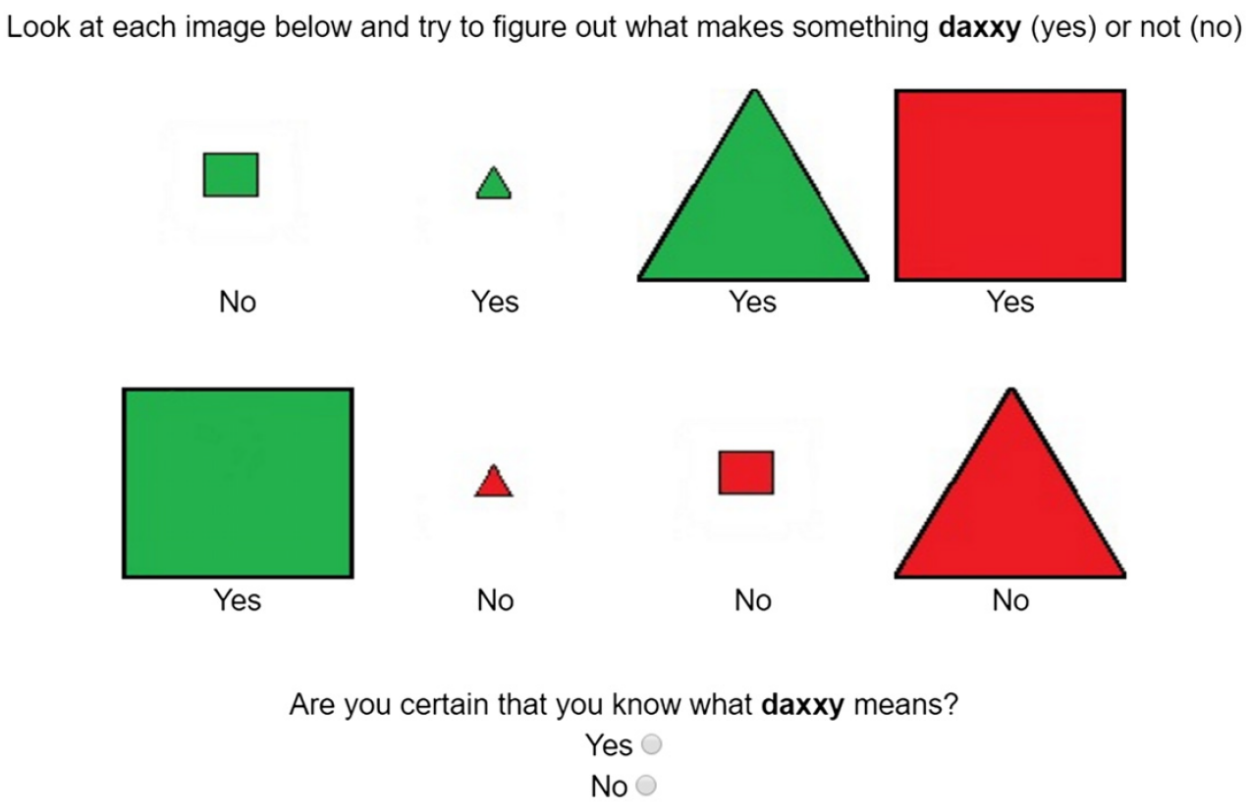

Figure 2-5: In Experiment 2, participants saw a single trial (as above), randomized between conditions.

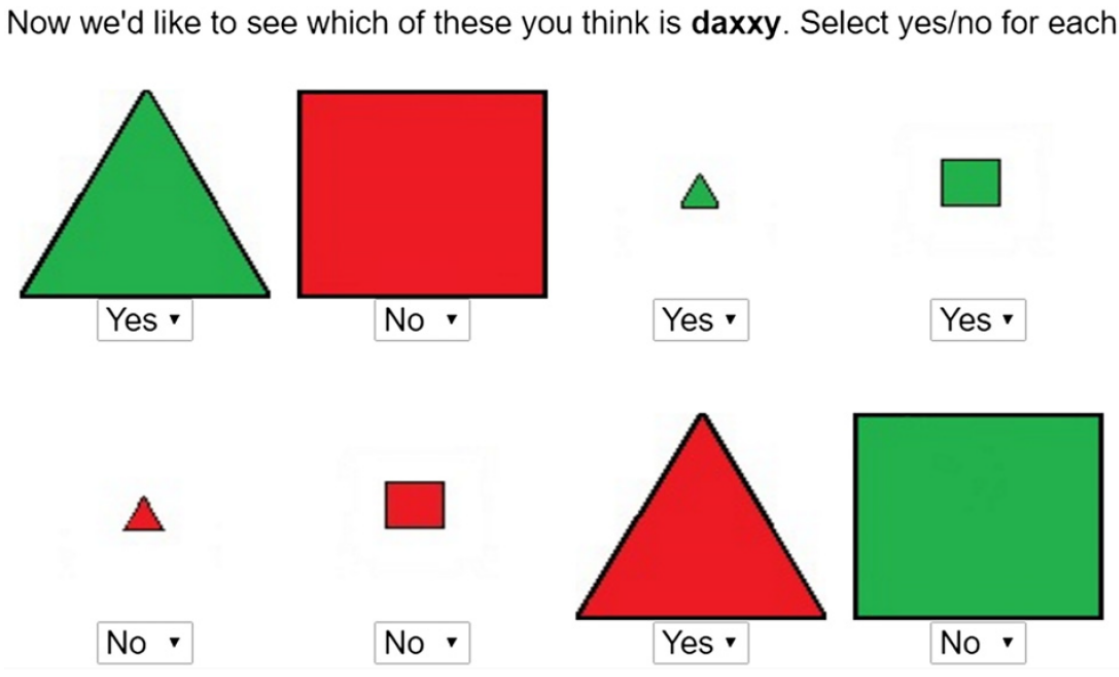

Figure 2-6: In Experiment 2, after responding regarding their certainty, participants labeled each stimulus to assess their accuracy.

Results

Unlike Experiment 1, accuracy was high across most conditions, with average accuracy ranging from 62% to 95% across conditions (see Figure 2-8 in the Appendices for details). This was likely due to participants viewing the data simultaneously and testing them immediately afterward. 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; see Figure 2-9 in the Appendices. 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 2-10 and Table 2-9 in the Appendices make clear that the results are similar to Experiment 1, with the best-performing predictors being behavioral measures. 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 nonobvious 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 dependent. 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 and 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.

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. We tested 536 participants on Amazon Mechanical Turk, using otherwise identical methods to Experiment 1 (see Appendix II-C for further details).

Results

Unsurprisingly, participant accuracies were similar to Experiment 1, replicating the general observed trends (see Figure 2-11 in the Appendices for details). 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; see Figure 2-12 in the Appendices).

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 (for details, see Figure 2-13 and Table 2-11 in the Appendices).

Discussion

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, perceptual certainty and certainty involving one’s memory of a fact (such as asking which country

has a higher population; Sanders et al., 2016) 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. One view is that certainty's function is as a guide to inform our beliefs and decisions. If certainty was fulfilling this 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 likely to be a useful or important cause of behavior and 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 et al., 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 predicted eyewitness accuracy, however, confidence judgments after memory "contamination" has occurred were no longer reliable (Wixted et al., 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 use of mixed-effect models. For example, individual differences in attention that influence performance at the 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 participants 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. Instead, a large component of certainty could reflect factors that are largely removed from the veridical probabilities that any given hypothesis is correct.

Acknowledgements

We thank the Jacobs Foundation, the Google Faculty Research Awards Program, and the National Science Foundation Research Traineeship Program (Grant 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.

Funding Information

Celeste Kidd, Jacobs Foundation (DE); Celeste Kidd, Google (http://dx.doi.org/10.13039/100006785); Louis Martí, National Science Foundation (http://dx.doi.org/10.13039/100000001), Award ID: 1449828.

Author Contributions

LM: Conceptualization: Lead; Data curation: Lead; Formal analysis: Lead; Investigation: Lead; Methodology: Lead; Project administration: Lead; Software: Lead; Validation: Lead; Visualization: Lead; Writing—original draft: Lead; Writing—review & editing: Lead. FM: Formal analysis: Supporting; Methodology: Supporting; Writing—review & editing: Supporting. SP: Conceptualization: Supporting; Formal analysis: Supporting; Methodology: Supporting; Resources: Supporting; Supervision: Equal; Writing—review & editing: Supporting; Validation: Supporting. CK: Formal analysis: Supporting; Funding acquisition: Lead; Methodology: Supporting; Resources: Conceptualization: Supporting; Lead; Supervision: Equal; Writing—review & editing: Supporting; Validation: Supporting.

III. Latent Diversity in Human Concepts

Louis Martí, Shengyi Wu, Steven Piantadosi, & Celeste Kidd

Many social and legal conflicts come down to differences in semantics. Yet, semantic variation between individuals and people's awareness of this variation have been relatively neglected by experimental psychology. Here, across two experiments, we quantify the amount of agreement and disagreement between ordinary semantic concepts in the population, as well as people's meta-cognitive awareness of these differences. We collect similarity ratings and feature judgements, and analyze them using a non-parametric clustering scheme with an ecological statistical estimator to infer the number of different meanings for the same word that is present in the population. We find that typically at least ten to twenty variants of meanings exist for even common nouns, but that people are unaware of this variation. Instead, people exhibit a strong bias to erroneously believe that other people share their particular semantics, pointing to one factor that likely interferes with political and social discourse.

Statement of Relevance

Cognitive science has long debated the degree to which common word meanings differ across individuals. Combining empirical data with state of the art modeling techniques, we statistically quantify the number of distinct concepts for 20 words across the population. We find strong evidence that the probability a single concept exists for each word is very small, and the most likely scenario is that roughly ten to twenty concepts exist, even for everyday nouns. These results suggest that fundamental conceptual differences at the lexical level extend to political and social discourse and underlie many semantic disagreements.

Introduction

Children learn word meanings through experience, and experiences differ between people. This suggests that even when two individuals use the same word, they may not agree precisely on its meaning. Indeed disagreements about meaning can be found in debates about the meaning of terms like "species" (Zachos, 2016), "genes" (Stotz et al., 2004), or "life" (Trifonov, 2011) in biology; "curiosity" (Grossnickle, 2016), "knowledge" (Lehrer, 2018), or "intelligence" (Sternberg, 2005) in psychology; and "measurement" in physics (Wigner, 1995). Ernest Mach and Albert Einstein even disagreed about what constitutes a "fact" (de Waal & ten Hagen, 2020); in contemporary society, social issues hinge on the precise meaning of terms like "equity" (Benjamin, 2019), "pornography" (Jacobellis v. Ohio, 1964), "peace" (Leshem & Halperin, 2020), or the "right to bear arms" (Winkler, 2011). Such debates often end up in the legal system. For example, in 1893, the U.S. Supreme court decided in Nix v. Hedden that, for tax

purposes, a tomato counted as a vegetable, not a fruit, stating that the law followed the "ordinary meaning" (see Goldfarb, 2021) of words rather than their botanical meaning.

Despite the frequency with which word meanings are debated, there have been few efforts to understand and quantify such variation in mental representations using tools of cognitive psychology. In the 1970s, Labov (1973) examined individual differences in people's classifications in a simple two-dimensional space of stimuli, cups and bowls that varied in height and width. This work found people often disagreed in atypical cases, a finding that holds in other domains (McCloskey & Glucksberg, 1978). Psychometrics developed multidimensional scaling methods (Torgerson, 1952; Shepard, 1962; Shepard, 1980) which account for individual differences, starting in the 1960s (Tucker & Messick, 1963; McGee, 1968; Carroll & Chang, 1970; Bush, 1973; Takane et al., 1977; Bocci & Vichi, 2011), with recent implementations (Okada & Lee, 2016) providing the advantages of generative Bayesian statistical inference (Gelman 2013; Kruschke, 2010). However, these tools have not been used to study population variation itself. Variation has been documented with amount of training or specialization—for instance, philosophers view "knowledge" differently than other academics and non-academics (Starmans & Friedman, 2020), and specialists often develop a specialized lexicon (Clark, 1998). Relatedly, recent results show that adults (Shtulman et al., 2020) and children (Sumner et al., 2019) will sometimes use words without real understanding.

While prior work has cataloged individual differences in typicality judgements using various natural and artificial categories (Barsalou, 1987; Verheyen & Storms, 2013; Koriat & Sorka, 2015; Hampton & Passanisi, 2016), no one has sought to robustly quantify how many varieties of ordinary concepts exist in the population. A primary challenge is that there are no complete accounts of human conceptual representation (see, e.g. Laurence & Margolis, 1999) and therefore people's representations must be measured indirectly. The approach of probing conceptual representations via linguistic labels has a long and fruitful history (Rosch & Lloyd, 1978, Lupyan & Thompson-Schill, 2012). In line with this tradition, we ran two experiments, collecting people's judgements of similarity between concepts and judgements of conceptual features respectively. The similarity experiment asked people to judge whether, for example, a penguin is more similar to a chicken or a whale. Similarities have been seen as foundational to some aspects of meaning, both in classic work (Shepard, 1962; Shepard, 1980; Shepard 1962; Barsalou 1989) and in more recent semantic representational theories (Landauer & Dumais, 1997; Mikolov et al., 2013). The feature experiment first freely elicited features, and then asked a group of participants to rate those on each concept. For example, participants judged whether a penguin was "majestic". We gathered ratings in two domains: common animals and politicians. These domains allow us to characterize diversity for high-frequency nouns, which may be most likely to be shared. We contrast this with politicians, which might vary among individuals with distinct political beliefs, as do concepts and language concerning morality (Graham et al., 2009; Frimer, 2020). We note that similarity judgements and features have well-known limitations, including for example that similarities are sensitive to the respects with which similarity is computed (Tversky & Gati, 1978; Medin et al., 1993; Gentner & Markman 1997; Markman & Gentner, 1993). However, because we are interested in quantifying diversity, it is not important to the experiment that such features and similarities do not completely

characterize people's conceptual knowledge. Differences in features and similarities still indicate that there are some underlying conceptual differences.

Importantly, we asked participants to make the same similarity ratings and feature judgements multiple times, allowing us to determine how reliable and stable the ratings were. This was important for our primary analysis because it allowed us to include in the modeling the possibility that people actually had consistent concepts, but they were just noisily measured, which could make it look as though people had different conceptual representations when in fact they did not. Our main results showing multiple concepts in the population therefore reflect statistical evidence of multiple concepts above and beyond response inconsistencies. Our primary analysis uses a non-parametric Bayesian clustering model in order to infer how many types of each concept (clusters) were likely to be present in our sample—for example, based on similarity judgements, how many different concepts of "finch" did people exhibit? This clustering method does not presuppose a fixed number of clusters, but infers a distribution of how many clusters are likely present based on the data. The distribution on the number of concepts combines two competing pressures: on the one hand, we should be biased to prefer a small number of clusters since this is a simpler theory. In the absence of data, the number of clusters should not be "multiplied without necessity", in the words of Ockham's Razor. Simultaneously, we should prefer a clustering which does a good job of explaining the data. Here, that means that the inferred clustering should predict responses, meaning that two individuals in the same cluster should give similar responses (see Figure 3-1). We use a non-parametric scheme (Gershman & Blei, 2012; Anderson, 1991; Pitman, 1995) which translates both of these pressures into probability theory, and then balances—optimally, in a precise sense–between the two. In particular, this clustering approach (see Methods) infers a probabilistic assignment of participants to latent "concept" clusters such that there are as few clusters as possible, while still being able to adequately explain their response patterns. This inference critically depends on the reliability of subject responses and only using this model are we able to infer the number of clusters that likely generated the data.

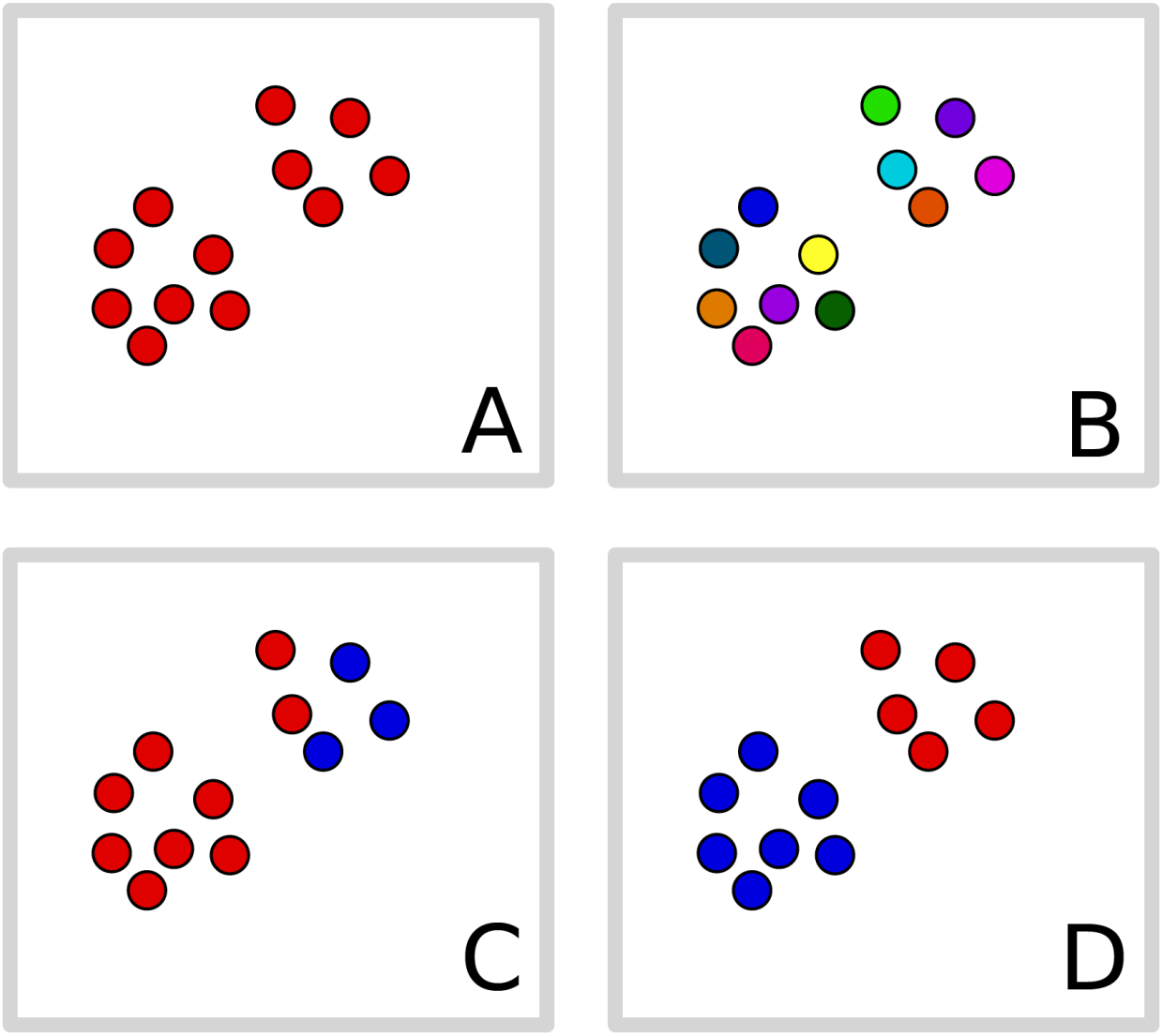

Figure 3-1: Hypothetical clustering of response vectors, here visualized in 2D. The simplest solution is to put all points into the same cluster (A), but then responses (locations) are not well-explained by clusters. If each point is in a separate cluster (B) then each point is perfectly predicted by the cluster, but the solution is complex. A compromise like (D) finds a small number of clusters that adequately explain the data. The correct clustering (D) will be preferred over alternatives even with the same number of total clusters which fit the data less well (C).

However, we are also interested in the number of clusters present in the population beyond our experimental sample. To quantify this, we used an estimator from ecology (Chao & Chiu, 2016). This model is more commonly used in species estimation in population ecology, where one might sample animals, observe how many of each species were collected, and estimate the total number of species present in the world (i.e. outside of the sample) from the distribution in the sample; closely related techniques can be found in (Good, 1953; Gale & Sampson, 1995). Here, we use the

most likely clustering of individuals to estimate the total number of concepts present in the world, outside of our sample.

Finally, we quantified people's metacognitive awareness of differences by asking participants to report what proportion of other people they expected to agree with them about their similarity judgements, and compared these reports to the observed agreement levels.

These methods allow us to test a variety of novel hypotheses about variation in human conceptual systems. First, by examining the estimated number of clusters (both in the sample and the general population), we evaluate how many measurably distinct representations can be found in the population. We note that this estimate is necessarily conservative since it is derived by similarities to a relatively small number of other nouns; larger and more detailed experiments might reveal more conceptual variation. Despite this conservativity, our results indicate that there is substantial variation present, even for these common nouns. Because this inference relies on a probabilistic model which incorporates multiple-measurement reliability, these clusters cannot be due to measurement noise. Our results also indicate common nouns and politicians have roughly the same number of different concepts in the population: both reflect substantial diversity. Finally, the meta-cognitive results show that people are generally unaware of these differences: most people expect that most others will answer the same way that they do. This lack of awareness suggests that such latent variation in what words are thought to mean may underlie disagreement on broader social and political issues.

Results

Experiment 1

We recruited 1,799 participants on Amazon Mechanical Turk. Half were asked to make similarity judgements about animals (finch, robin, chicken, eagle, ostrich, penguin, salmon, seal, dolphin, whale) and the other half to make judgements about U.S. politicians (Abraham Lincoln, Barack Obama, Bernie Sanders, Donald Trump, Elizabeth Warren, George W. Bush, Hillary Clinton, Joe Biden, Richard Nixon, Ronald Reagan). Each participant was randomly assigned to a single target from one domain (e.g. "finch"), presented with 36 unique pairs of other objects in the domain (drawing from the 10 objects in each domain), and asked which was more similar to the target. Thus, participants responded to queries such as "Which is more similar to a finch, a whale or a penguin?" Each trial was shown twice (for a total of 72 trials) in order to measure response reliability (calculated as the percentage of trial-pairs with identical responses) and detect trial-by-trial idiosyncratic features of stimuli. To quantify metacognitive awareness of diversity, participants were also simultaneously asked on every trial to guess how many people out of 100 would agree with their response.

We coded each participant's responses to a single word as a binary vector, corresponding to the forced-choice similarity rating between every other pair of items. In modeling, we assumed that individual vectors were sampled from a collection of latent clusters that specified an average response vector. We used a nonparametric Bayesian technique, a Chinese Restaurant Process (Gershman & Blei, 2012; Anderson, 1991; Pitman, 1995), to model a posterior distribution on the number of clusters for each word

independently, marginalizing out the average response vectors for each cluster and assuming a reliability given by the overall average reliability. We note that this clustering model works in the space of response vectors, not in the lower-dimensional psychological space itself; thus, our approach does not explicitly model correlations that may exist between items, but also does not require us to make assumptions about the dimensionality or metric properties of the latent space. This technique permits us to find a distribution over the number of clusters present in the population, taking into account both the reliability of individual responses and uncertainty about the latent response vector characterizing each cluster (e.g. what each participant's "finch" cluster corresponds to in terms of similarities). The model builds in a prior preference for fewer clusters—a version of Ockham's Razor—but we also present results with no such prior. The maximum a posteriori clusterings found in sampling were additionally put through a species-count estimator which estimates the true number of clusters present in the global population, beyond our finite sample size (Chao & Chiu, 2016). This estimator uses sampled individuals which are observed to fall into a distribution of species and estimates the total number of species (here, clusters) in the population at large.

The overall subject reliability is around 90% (see Figure 3-6 in Materials and Methods), indicating subjects are both not responding with random guesses, nor are they responding with ad hoc responses that vary from trial to trial. Subject responses likely reflect stable aspects of how they conceptualize these concepts throughout the context of the experiment.

Figure 3-2 shows a visualization of participants' similarity judgements using distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008). This technique places individual participants' response vectors in a 2D plane such that nearby participants give similar response vectors. The closer two points are together, the more closely their concepts align; however, these scales are relative and cannot easily be compared across plots. Points in this plot have been colored according to the maximum a posteriori assignment of participants to clusters according to the clustering model, which was run independently from t-SNE. This figure illustrates that two independent methods provide convergent characterizations of how people are distributed in the space since each color (generated according to the clustering model) tends to be in a single spatial position (generated by t-SNE). Note that the color assignments do not perfectly match spatial arrangements, likely due to t-SNE dimensionality reduction and different trade-offs being applied to edge-case participants by our algorithm and t-SNE.

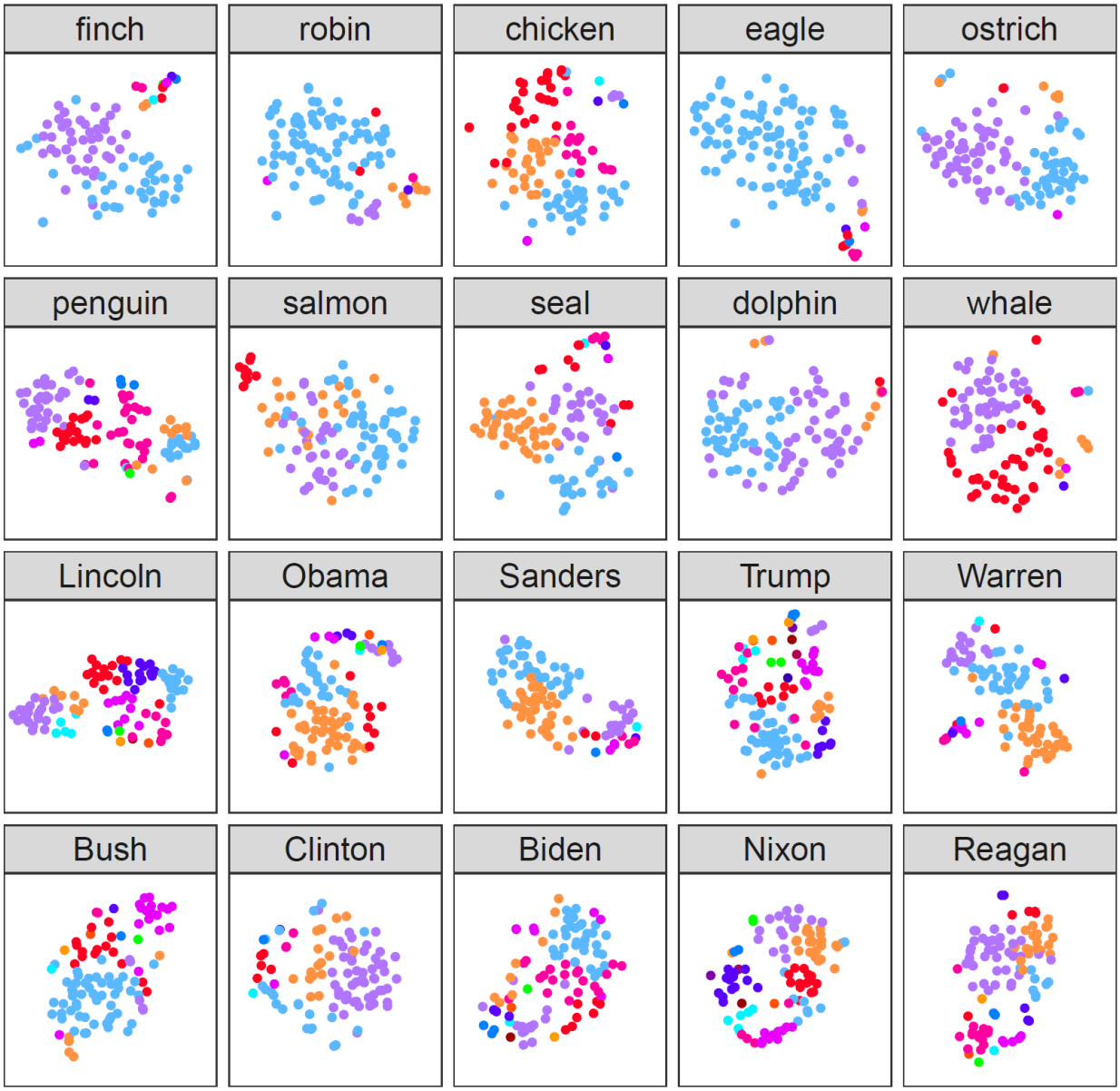

Figure 3-2: Distances between participants' conceptual representations of each target concept using distributed stochastic neighbor embedding. In this visualization, the distances between two points approximate the distance between their full rating vectors. Each plot is on the same scale. Additionally, each data-point is colored with the cluster they were assigned to in our clustering analysis, showing that the t-SNE clustering finds similar groupings.

To understand the number of concepts in the population, we first look at the posterior distribution over the number of clusters inferred. Figure 3-3 shows the estimated number of conceptual kinds (y-axis) for each semantic domain (subplot), as a function of the number of participants included (x-axis). This figure shows that as our sample size increases from 10 to 100 individuals per concept, the number of estimated concepts reaches 7 to 12 for politicians and 4 to 10 for animals. The maximum a

posteriori clustering (in purple) and the ecological estimator (in blue) are in the range of 10-20 latent concepts in the population, and are higher for politicians than for animals. We find similar ranges even if we use a prior which is uniform on the clusterings (orange).

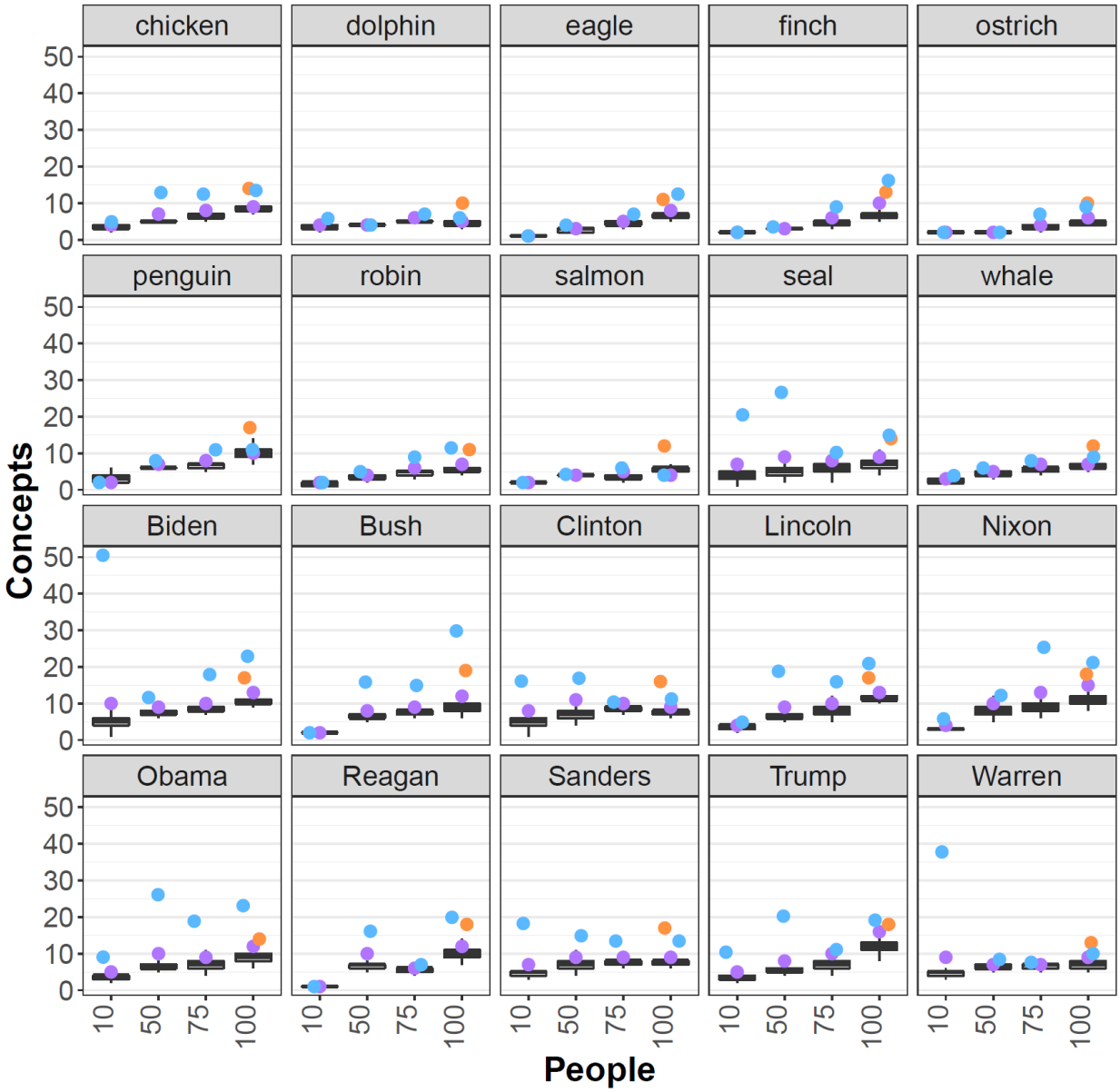

Figure 3-3: Estimated number of concepts (y-axis) depending on the number of people sampled (x-axis). Boxes show the median 50% quantiles of the number of unique concepts. Purple data points are the number of clusters for the maximum a posteriori clustering. Orange data points are the number of clusters for the MAP clustering with a uniform prior. Blue data points are a lower bound on the number of concepts estimated by the ecological estimator using the MAP clustering.

We note that the number of inferred concepts is not necessarily monotonically increasing in the number subjects, since additional subjects may shape the geometry of the space (e.g. providing evidence that two separate clusters are actually one wider cluster). In addition, most of the latent diversity can be found in small numbers of subjects–even distinct clusters can be found when examining 50 individuals. The point at which each subplot levels off is due to a combination of the reliability of individual responses, the number of items we sampled (sampling less results in fewer concepts), and the true number of concepts in the population. However, limited reliability and a finite number of items mean that our analysis is likely to under-estimate the number of clusters.

Figure 3-4 shows the probability that the population contains only one concept for each word, according to the clustering model (Recall that due to the limitations of the similarity measure, this is an overestimate). Political words are far less likely to have a single meaning than animal words, matching the patterns in the number of clusters in Figure 3-3. Generally, this provides strong statistical support to the idea that there are multiple meanings in the population for these terms, despite the fact that these multiple concepts all have the same word. However, if the distribution of participants to meanings tends to be heavily skewed (e.g. most participants have the same meaning), then this diversity might be relatively inconsequential. Figure 3-4 shows the probability that two randomly chosen individuals will have the same concept in this analysis, which is a relatively robust statistic since it depends largely on the frequency of the most common concepts for each word rather than the tails of the distribution. This agreement probability averages to around 25-50% for animals and 10-30% for politicians. This indicates that most individuals one encounters will tend to have a measurably different conceptual representation. Again, this is likely to overestimate the true rate of agreement since we only tested a small number of questions.

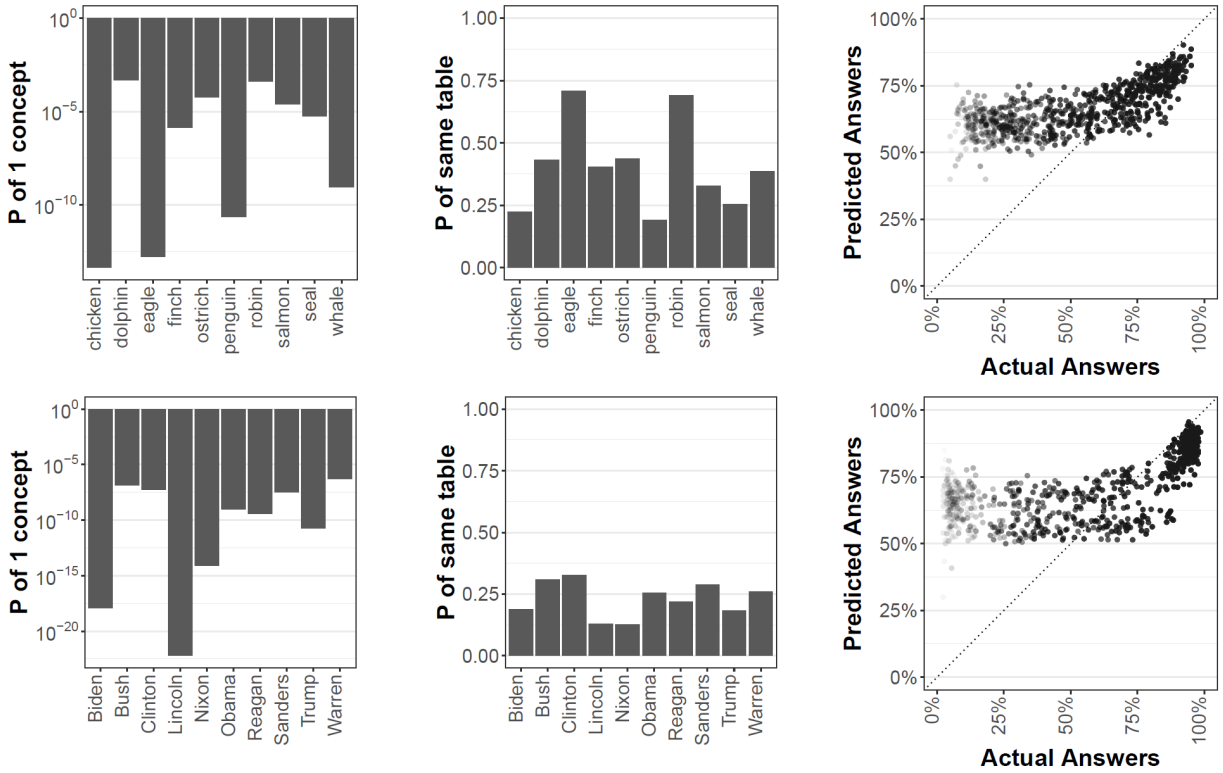

Figure 3-4: Left: Probabilities that only a single conceptual representation for each word exists, with log-axis, showing essentially zero probability for any word. Middle: Probability that two random individuals will share the same table (i.e. concept), showing generally low rates of agreement. Right: Predicted answers (y-axis) vs. actual answers (x-axis), showing people tend to overestimate others' rate of agreement compared to the truth (line y = x).

Most importantly, our results show that people are generally not aware of these differences. Figure 3-4 shows the agreement rate on responses (x-axis) compared to people's predicted estimates of agreement (y-axis). If people understood the population's variation in responses, the trials shown in this plot would all fall along the y = x line. Instead, this figure shows that for most of the range of actual agreement (e.g. ~ 0% - 80%) people tend to consistently believe that about 2/3 of participants will agree with them, no matter what true proportion actually do. This is true even for the lowest agreement responses: most participants believe their response is in the majority even when essentially 0% of other participants agree with them. This is unlikely to be due to a failure to engage this aspect of the task because participants do reliably increase their estimates on the highest agreement items (e.g. ~ 80% - 100%), which results in a reliable rank-order correlation overall (Spearman's $\rho = 0.45$, $p < 0.001$). The increase, though, is not well-calibrated to the population variation. Moreover, these patterns likely reflect meta-cognitive limitations (Gopnik et al., 1997; Wimmer & Perner, 1983; Goldman et al., 2006) rather than differences in effort or motivation because these trials were interspersed with the main task, which had very high within-subject reliability.

Experiment 2

Experiment 2 consisted of two parts: feature elicitation and feature rating. In feature elicitation, we recruited 16 registered users on Prolific. Half of the participants were asked to list 10 single-word adjective features for each of the 10 animals in Experiment 1. The other half were asked to list 10 single-word adjective features for each of the 10 U.S. politicians in Experiment 1. We kept all features that were mentioned more than once after removing non-adjectives, inappropriate words, and typos, as well as combining synonyms.

Then, 1,000 registered users on Prolific were asked to rate either 105 animal features or 105 politician features from the feature elicitation experiments. Each participant was randomly assigned to rate features of two animals (e.g. “dolphin” and “whale”) or two U.S. politicians (e.g. “George W. Bush” and “Hillary Clinton”). Participants were asked questions such as “Is a finch smart?” and responded by clicking either the “Yes” or “No” button on the screen. Each question was asked twice to measure response reliability. Thus, each participant saw 420 question trials.

Participant reliability was high with an average reliability of 86% during feature rating, indicating participants were not responding with random guesses. Similar to Experiment 1, subject responses likely reflect stable aspects of subjects’ conceptual representations.

Clustering participants based on their feature ratings serves as a conceptual replication of Experiment 1. The number of concepts found is 6 to 11 for politicians 5 to 8 for animals, compared to 7 to 12 for politicians and 4 to 10 for animals in Experiment 1 (see Figure 3-10 in the Appendices). Likewise, the ecological estimator results in 8 to 30 latent concepts in the population, compared to 10 to 20 in Experiment 1. Such similar findings, despite a very different paradigm, provides convergent support for conceptual diversity.

Figure 3-5 shows mean agreement for a sample of features and concepts. Many features show near universal agreement among participants. A similar number of features show large disagreement among participants. Most participants agree that seals are not feathered but are slippery while disagreeing as to whether they are graceful. Likewise, most participants agree that Trump is not humble and is rich, but there is high disagreement as to whether he is interesting. These sorts of disagreements reflect the different conceptual representations possessed by our participants.

Figure 3-5: A sampling of feature responses for 3 animals and 3 politicians (y-axis). The x-axis plots the mean percentage of “yes" responses for a given feature. Features on the left are generally agreed to not apply and features on the right are usually agreed to be applicable. Features in the center however, show high disagreement among participants and are the primary features responsible for differing conceptual representations among participants.

Discussion

We report statistical evidence of more than one variant of concepts in the population. In fact, we find that most people the average language user meets will not share their same concept. These results are unexpected in part because the measures we used are coarse. If one could gather an arbitrary amount of data, one might expect to find small differences between people: one interlocutor might have specific memories that make their representation idiosyncratic, perhaps different from anyone else. However, our experimental approach was based on judging similarities and features—not an exhaustive inventory of each person’s memories or associations—and we were nonetheless able to statistically justify measurably distinct representations,

even for common nouns. If differences can be detected with these methods, it indicates that there is substantial variation in the population for lexical meanings. This variation exists despite the fact that people use the same word for each concept, and people are relatively unaware that others will tend to give differing similarity judgements.

However, our results do not support the notion that every single use of a concept is distinct or entirely idiosyncratic (Casasanto & Lupyan, 2015): subjects did group into clusters and did provide highly reliable responses across trials. We emphasize, though, that studies with more items, or items that focus more on corner cases, might find greater diversity than reported here. Moreover, the subject pool in our experiment was relatively homogeneous, and future studies of cultural differences may point to more diversity in word usage based on diversity of experience (Clark, 1998). Indeed, while our method allows us to quantify conceptual diversity, it does not pinpoint what specific representational differences drive this diversity. These differences may indeed go deep with respect to theories and interrelations between the concepts studied and others (Murphy & Medin, 1985; Medin & Rips, 2005; Gelman & Legare, 2011).

In general, theories of word learning and conceptual development will need to work out how human language users acquire distinct representations for shared words. In turn, theories of communication and language use (e.g. Wilson & Sperber, 2002; Grice, 1989) will need to address both differences in word referents, and lack of awareness of those differences. People's general obliviousness to variation has important implications for productive discourse structure, and has been studied by psychologists in more general forms such as the false consensus effect (Marks & Miller, 1987) and egocentric bias (Ross & Sicoly, 1979). Fundamental misunderstandings may originate with individuals using the same word for distinct conceptual representations or under different contexts. Indeed, such differences in word meanings might underlie many classic philosophical questions (Piantadosi, 2015). Generally, our results may help to explain why "talking past each other" appears to be common in social and political debates: the common ground of even the most basic word meanings is only imperfectly shared.

Materials and Methods

Experiment 1 was run using a custom built web interface on Amazon Mechanical Turk on 8/20/19 through 8/22/19 (animals) and 9/11/19 through 9/12/19 (politicians). Participants were instructed to "decide which [animal/politician] is more similar to [target concept]" and "asked to guess how many people out of 10 would agree with you." All participants were required to be from the U.S. and have a minimum 95% approval rating from previous tasks. Experiment 2 was run on 04/23/21 through 05/09/21 (animals) and 05/13/21 through 05/17/21 (politicians) using Prolific and Qualtrics. Participants were all above 18 years old, fluent English speakers, and physically present in the United States based on pre-screening questions. Responses were recorded on a secure server and no participants were excluded from data analysis. All participants were paid at a rate of $10 an hour. This study was approved by the Committee for Protection of Human Subjects at University of California, Berkeley (CPHS # 2018-12-11675). Informed consent was obtained from all subjects. All methods were performed in accordance with relevant guidelines and regulations.

Clustering Methods

Responses were clustered using a non-parametric, Bayesian clustering model, a "Chinese restaurant process" (51), with a custom implementation in Python. If $x = \{x_1, x_2, ..., x_k\}$ denotes the number of subjects in each cluster (for a given word), and n denotes the total number of subjects, this model assigns x, a partition on individuals, a prior of

$$\frac{1}{n!}\prod_{i=1}^{k}(x_i - 1)!.$$

Within each cluster, we use a Beta-Bernoulli likelihood where subjects assigned the same cluster are assumed to generate the same latent vector of answer probabilities. We marginalized out each cluster's probability vector. Thus, if aij and bij are the number of each type of response to question j in cluster i, then the marginal likelihood of those responses is,

$$\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \cdot \frac{1}{\Gamma(x_i+2\alpha)} \prod_{j=1}^{x_i} \frac{\Gamma(a_{ij}+\alpha)}{\Gamma(a_{ij}+1)} \cdot \frac{\Gamma(b_{ij}+\alpha)}{\Gamma(b_{ij}+1)}.$$

Here, α characterizes the noise level assumed by the likelihood. We set the single likelihood parameter α = 0.16 such that two samples from a Bernoulli with parameter p ~ Beta(α, α) agreed with each other with probability 0.88, which is the proportion of time subjects' second and first responses agreed (analysis of the dependence of the results to the assumed α is in the Appendices).

Inference was run using a Gibbs sampler, using both the prior (1) above and a uniform prior. All runs used the same likelihood (2). The Gibbs sampler followed standard inference techniques by selecting each individual and resampling a cluster assignment for them based on the posterior probability of each possible cluster assignment (including assigning to a new cluster). The sampler was initialized with a configuration where each individual started in the same cluster. These sampling methods require iterations of burnin before they converge to a stable set posterior distribution. We assessed convergence using multiple runs and ensured that chains arrived at the same solution. Figure 3-7 in the Appendices shows the convergence of three chains for each concept over 300k iterations. Our final run used a 100k iterations of burn-in and an additional 200k iterations of sampling.

Ecological Estimator

Finally, we use an ecological estimator from (Chao & Chiu, 2016), extending a previous estimator (Colwell & Coddington, 1994), in order to approximate the total number of concepts in the population. This estimator uses the total number of observed clusters (concepts) and the total number of sampled individuals in order to estimate how

many concepts were likely unobserved. Like, for instance, Good-Turing estimation (Good, 1953), this estimation depends on the number of clusters containing a single person, but also includes additional terms. Let fi denote the number of clusters containing i individuals in the maximum a posteriori Bayesian clustering. The estimator is based on $\widehat{S}_{Chao1}$, given by,

$$\widehat{S}_{Chao1} = \begin{cases} S_{obs} + \frac{(n-1)}{n}\frac{f_1^2}{2f_2}, & if\ f_2 > 0 \\ S_{obs} + \frac{(n-1)}{n}\frac{f_1(f_1-1)}{2}, & if\ f_2 = 0. \end{cases}$$

Here, S_{obs} denotes the number of observed clusters and n is the number of participants sampled. The estimator we used adjusts $\widehat{S}_{Chao1}$ to yield $\widehat{S}_{iChao1}$,

$$\widehat{S}_{iChao1} = \widehat{S}_{Chao1} + \frac{(n-3)}{n} \cdot \frac{f_3}{4f_4} \cdot max(f_1 - \frac{(n-3)}{(n-1)}\frac{f_2 f_3}{2f_4},\ 0).$$

Acknowledgements

The authors thank the Kidd Lab and the Computation and Language Lab for feedback.

Funding

DARPA (Machine Common Sense TA1, BAA number HR001119S0005, CK & SP)
NSF (Division of Research on Learning, Grant 2000759, CK & SP)
Human Frontier Science Program (RGP0018/2016, CK)
Berkeley Center for New Media (CK)
The Jacobs Foundation (CK)
Google Faculty Research Awards in Human-Computer Interaction (CK)

Author contributions

Conceptualization: LM, CK, SP
Methodology: LM, CK, SP
Investigation: LM, SW
Software: LM, SP
Formal analysis: LM
Visualization: LM, SP, SW
Writing – original draft preparation: LM
Writing – review and editing: LM, CK, SP, SW
Supervision: CK, SP
Funding acquisition: CK, SP

IV. “Fringe” beliefs aren’t fringe

Louis Martí, Adam Conover, & Celeste Kidd

COVID-19 and the 2021 U.S. Capitol attacks have highlighted the potential dangers of pseudoscientific and conspiratorial belief adoption. Approaches to combating misinformed beliefs have tried to “pre-bunk" or “inoculate" people against misinformation adoption and have yielded only modest results. These approaches presume that some citizens may be more gullible than others and thus susceptible to multiple misinformed beliefs. We provide evidence of an alternative account: it’s simply too hard for all people to be accurate in all domains of belief, but most individuals are accurate most of the time. We collected data on a constellation of human beliefs across domains from more than 1,700 people on Amazon Mechanical Turk. We find misinformed beliefs to be broadly, but thinly, spread among the population. Further, although some beliefs are associated with others, we do not find evidence that individuals who adopt a single misinformed belief are more likely to engage in pseudoscientific or conspiratorial thinking across the board, in opposition to “slippery slope” notions of misinformation adoption.

Statement of Relevance

Psychological science has debated whether misinformed beliefs are localized within a relatively small subgroup of individuals, or whether these beliefs are dispersed across the population. Using a large sample size and demographic correction techniques, we demonstrate that misinformed beliefs are widely but weakly held, and that only some beliefs are slippery slopes into others.

Introduction

Recent events surrounding QAnon and COVID-19 conspiracies have highlighted the potential dangers of misinformation (Romer & Jamieson, 2020; Woko et al., 2020; Amarasingam & Argentino, 2020). In response, efforts to “pre-bunk” the conspiracies or “inoculate” the population against the spread of misinformation have arisen (Maertens et al., 2020; Pennycook & Rand, 2020; Roozenbeek & van der Linden, 2019), with thus far only modest results (Banas & Rains, 2010). The potential robustness of this approach depends upon an untested assumption of human psychology: that some individuals are more gullible than others, and that they can be made less gullible through training.

Previous studies of conspiratorial and pseudoscientific belief focus on populations who hold misinformed beliefs in a single domain, for example flat earthers (Landrum, 2021), anti-vaxxers (Martinez-Berman et al., 2020), climate change deniers (Uscinski et al., 2017), or incels (Young, 2019). However, examining beliefs in a single domain would necessarily make any variation appear as though some individuals are inherently more gullible than others. In fact, it is quite possible that everyone is trying their best to form beliefs that align with objective evidence in the world, and generally doing a decent but imperfect job across multiple knowledge domains.

We compare two hypotheses: one is that some individuals are fundamentally gullible and therefore are susceptible to multiple misinformed beliefs. The other is that it is simply hard to be accurate in all domains of belief. In that case, we would expect misinformed beliefs to be broadly, but thinly, spread among the population.

To understand how beliefs arise and spread, you must look at constellations of beliefs. Here, we do just that. We collect a large set of judgements on a host of different types of beliefs—including conspiracies, pseudoscience, and other non-evidence-based beliefs.[5] We use this data to understand the prevalence overall of many misinformed beliefs, as well as whether belief in one tends to predict belief in others, as is widely espoused in "slippery slope" arguments (Wood et al., 2012).

Our results demonstrate that rather than some portion of the population being gullible, most people hold one or more non-evidence-based beliefs. Our results suggest most of these beliefs do not predispose individuals to becoming more likely to adopt many other non-evidenced beliefs. The fact that misinformed beliefs are ubiquitous and generally not "gateway drugs" to other networks of misinformed beliefs has widespread implications for how we should structure efforts to combat misinformation in the world.

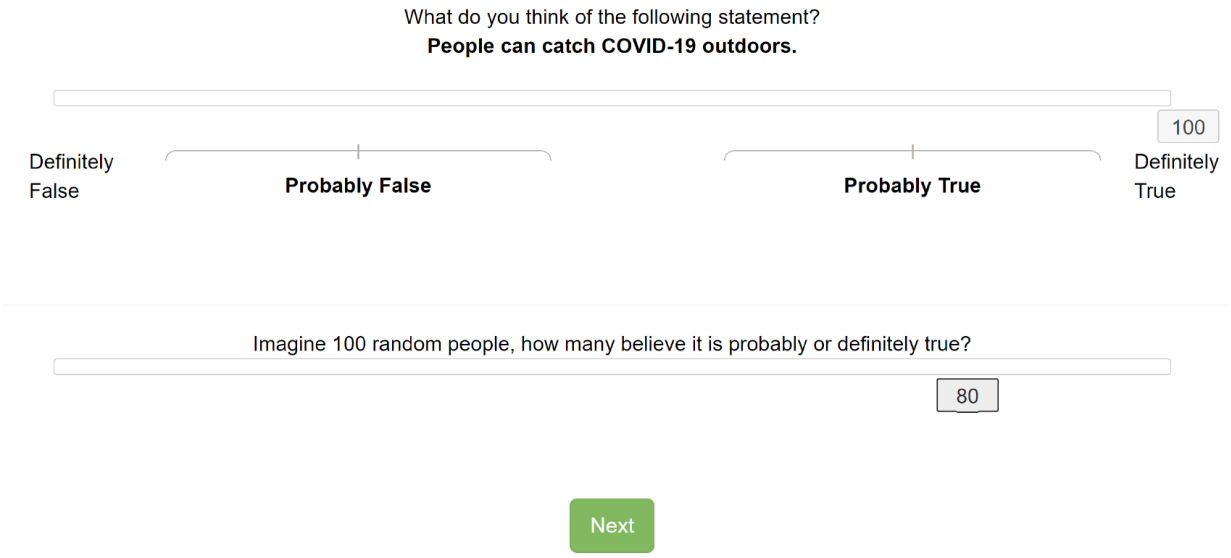

Figure 4-1: Participants saw 60 randomized trials as above.

[5] The misinformed beliefs we evaluated here included "conspiracies" and "pseudoscience", as well as other non-evidence-based beliefs, most but not all of which are commonly labeled "fringe beliefs". We recognize that these terms are not interchangeable, but investigate all to broadly understand how misinformed beliefs relate to one another in the population.

Results

Fringe beliefs are not fringe

We calculated a belief score for each participant representing the number of non-evidence-based statements that they believe are more likely than not.[6] While any given misinformed belief may be uncommon, when examined in aggregate, the vast majority of people believe in at least several misinformed statements (see Figure 4-2). In our sample, 98% of participants believe at least one statement[7], and 52% believe at least nine. It is important to note that this measure is conservative, as we only tested a small minority of all non-evidence-based beliefs. Interestingly, unsubstantiated COVID-19 beliefs are substantially less common than all other misinformed beliefs, possibly because they have had less time to spread throughout the population, or perhaps due to public health messaging (see Figure 4-5). The median General Pseudoscience score is 6 (out of a possible 18) while the median COVID Pseudoscience score is 3 (out of a possible 12).

[6] This score was reverse coded for the 14 statements that were factual.
[7] We excluded 319 participants. Remaining participant reliability was 88.7%. See Exclusion Criteria in Materials and Methods.

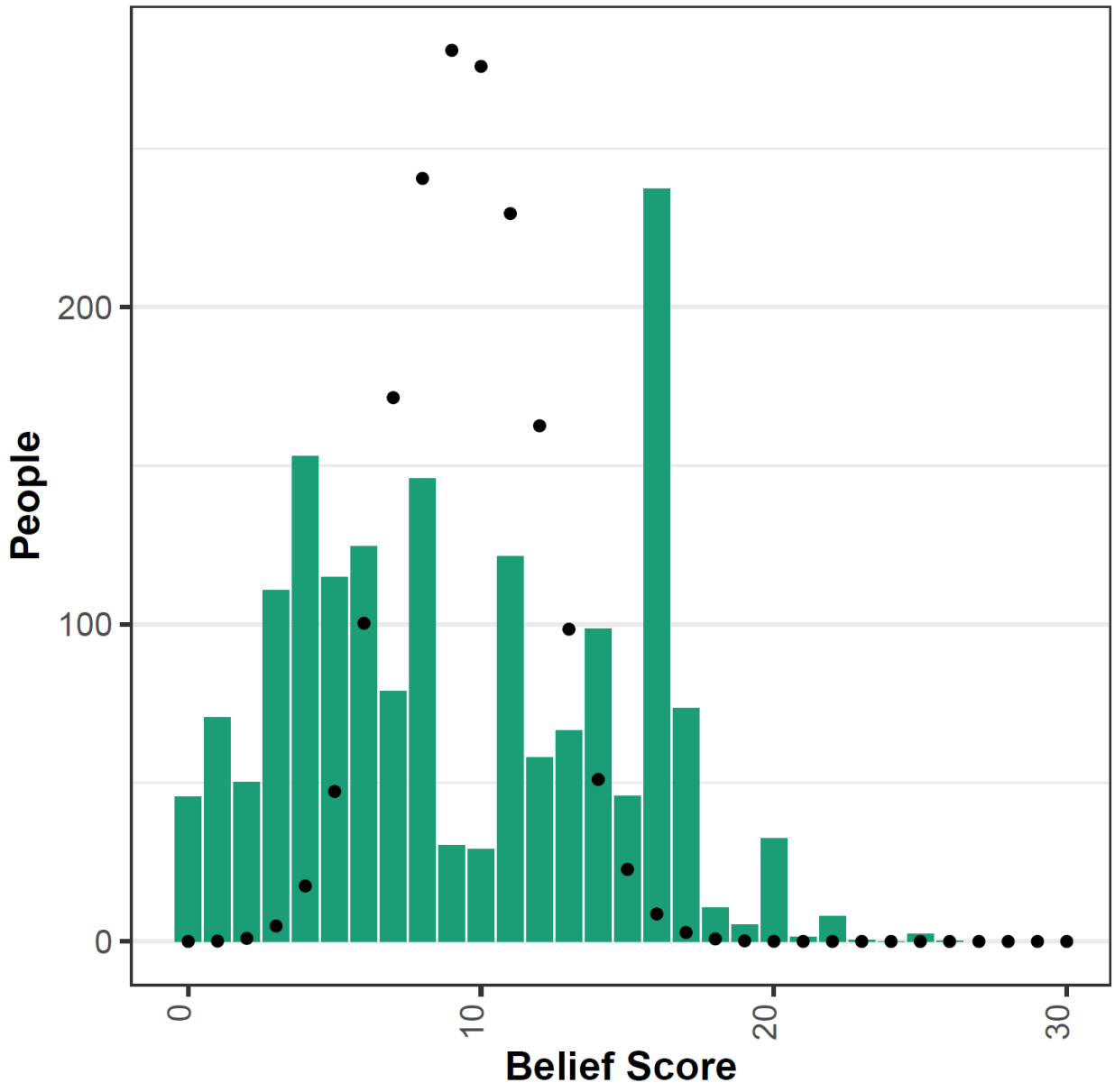

Figure 4-2: A histogram showing the number of people who believe x number of misinformed statements are more likely than not, after sampling bias correction. The maximum possible score is 30, one point for each statement. The median participant believes 9 out of 30 non-evidence-based statements. Despite being a conservative measure (we only tested a small minority of non-evidenced beliefs), we find that misinformed beliefs are widespread. Black dots represent the expected distribution of beliefs if all beliefs were independent of each other in a Poisson-binomial distribution. This analysis provides some evidence for both a traditional slippery slope, and a reverse slippery slope where individuals are less likely to believe other beliefs, as can be seen by green bars found above black dots.

Fringe beliefs are often weakly held

While these beliefs are widespread, they are not strongly held. Figure 4-6 shows the weighted aggregate of all responses across all sentences. Among all misinformed statements that participants believe are more likely than not, there still exists a high degree of uncertainty. Very few non-evidence-based statements are rated as 100% true, but the same level of uncertainty is not present when participants rule out a misinformed belief (an overwhelming number of responses rate them as 0% likely).

Only some beliefs are possible gateways

As Figure 4-7 shows, only some beliefs provide evidence for the slippery slope argument. Each boxplot partitions participants depending on whether or not they believe a particular misinformed statement. If a slippery slope existed, participants who believed any given misinformed statement would have a significantly higher conspiracy score than the participants who did not believe it. In other words, the boxes (which represent 95% confidence intervals) within a statement would not overlap and the conspiracy box would be higher. Instead, we find that this is only the case for 13 out of the 30 statements. As supported by a linear regression, if you believe that aliens are currently visiting the Earth, you are no more likely to believe other misinformed beliefs ($\beta = -0.0588$, $SE = 0.0048$, $t = -12.14$, $p < 0.001$). On the other hand, if you believe autism is caused by environmental toxins, you are more likely to believe other misinformed beliefs, as a linear regression confirms ($\beta = 0.1377$, $SE = 0.0036$, $t = 38.70$, $p < 0.001$). It is important to note that this is still not a confirmation of a slippery slope in these cases, as causality would need to be determined. Comparing COVID-19 scores in the same manner results in 7 out of 12 significant differences, indicating that perhaps the slope is a bit slipperier when dealing with misinformed beliefs which are closely related.

Figure 4-3 shows belief scores for participants based on randomly assigning each statement into one of two arbitrary groups. If beliefs were held completely independently the within subject correlation of both scores would be zero. Instead, we see a moderate correlation implying that while these beliefs are not deterministically held, they hold some influence over each other.

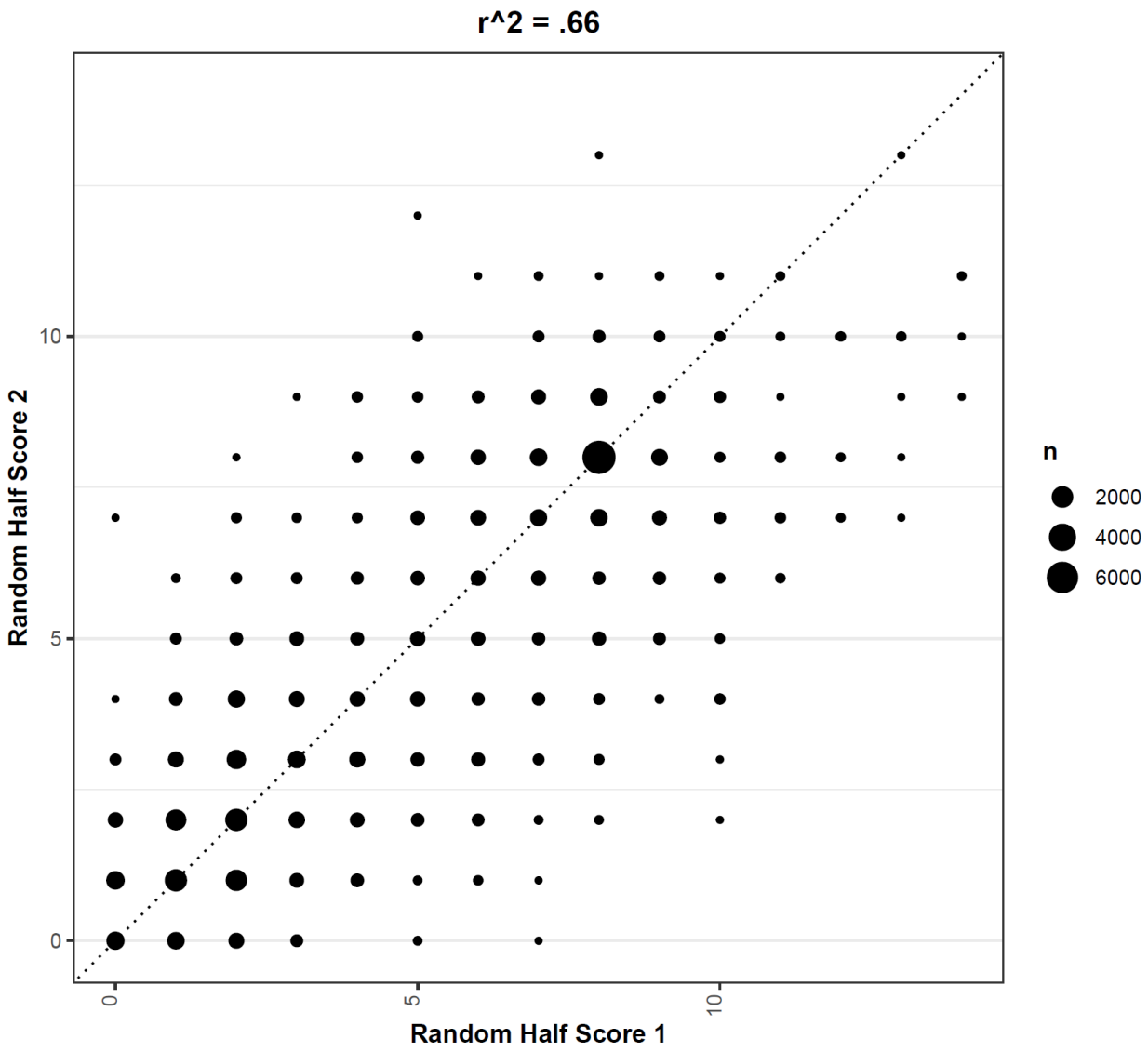

Figure 4-3: Belief scores for split statements. Statements were randomly split into two categories and participants were assigned belief scores for each category. If beliefs are held independently of each other we would expect no correlations between each score within participants. Instead, we find a moderate correlation with an R^2 of .66.

In aggregate, participants are very good at predicting beliefs

As Figure 4-4 shows, the predictions participants made about the prevalence of these beliefs were very accurate. With each data point representing a different statement, all of them are either just above or just below the line of perfect prediction y = x. A generalized logistic regression confirms this (β = 1.6201, SE = 0.0344, z = 47.12, p < 0.001).

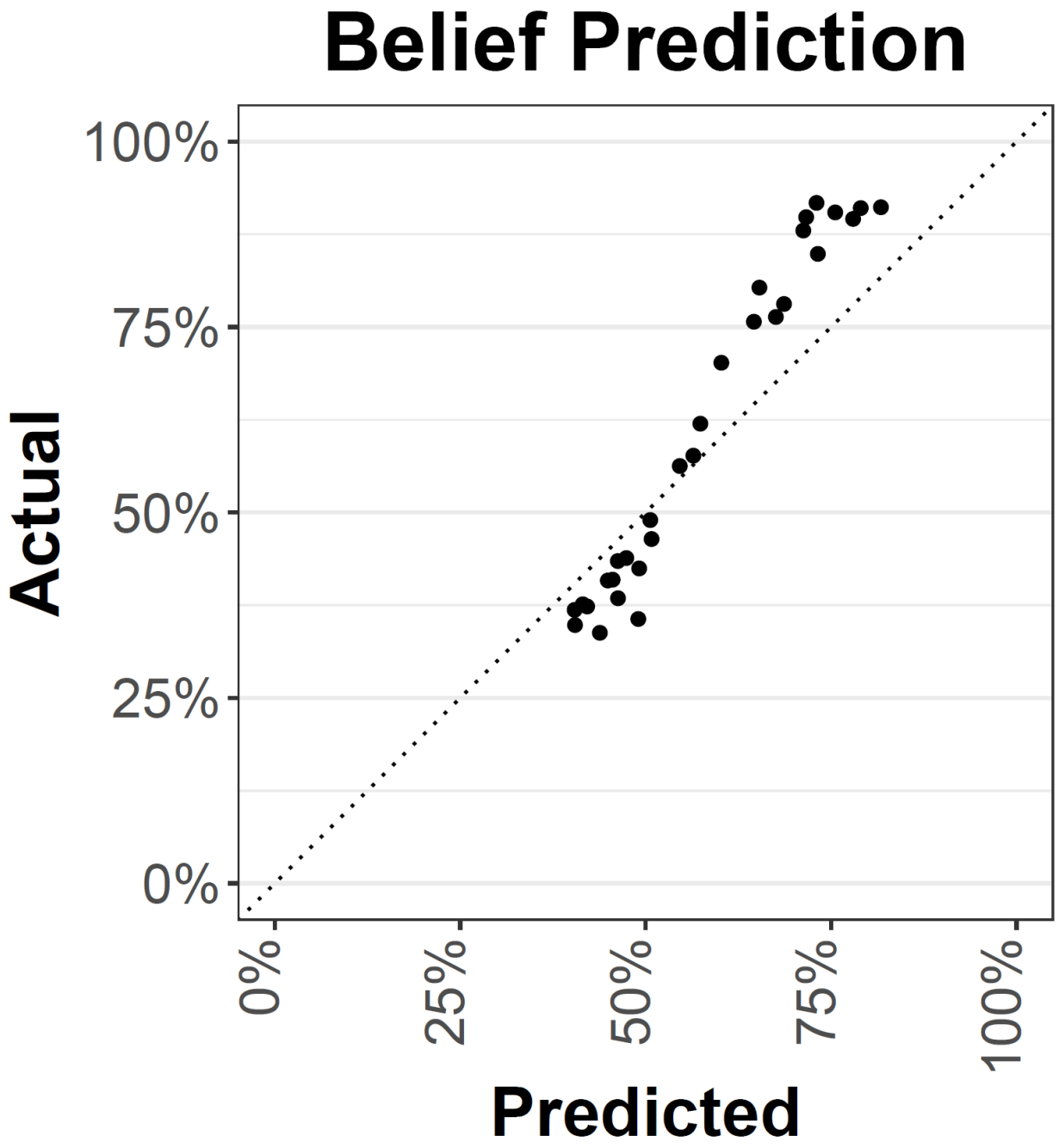

Figure 4-4: Mean predictions and the actual percentage of participants who believe each statement is more likely than not. Perfect predictions would lie along the y = x line; participants are very good at predicting the beliefs of others in aggregate.

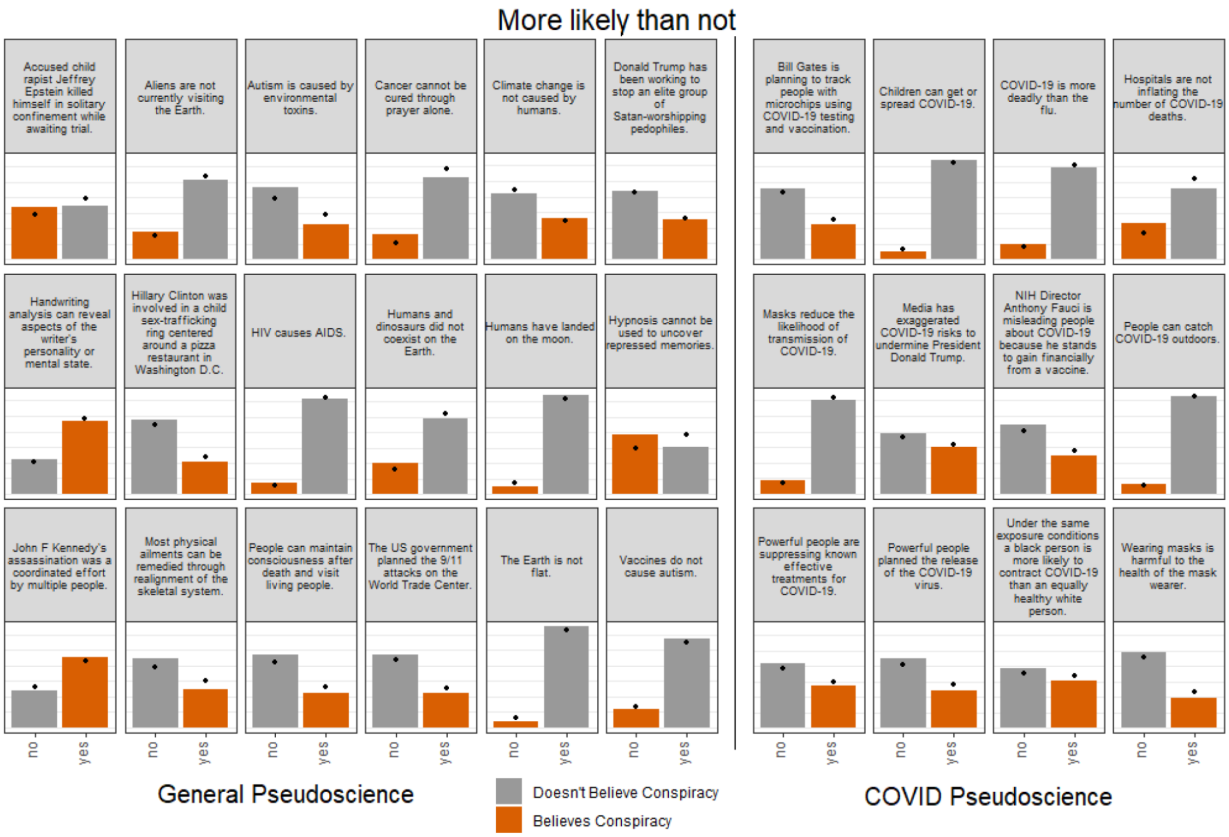

Figure 4-5: Proportions of people who believe each statement is more likely than not. Orange bars represent individuals who endorsed a misinformed statement while gray bars represent a rejection. Bars show data after sampling bias correction while the black dots show data before the correction. Misinformed beliefs are common and in certain cases, represent the majority view.

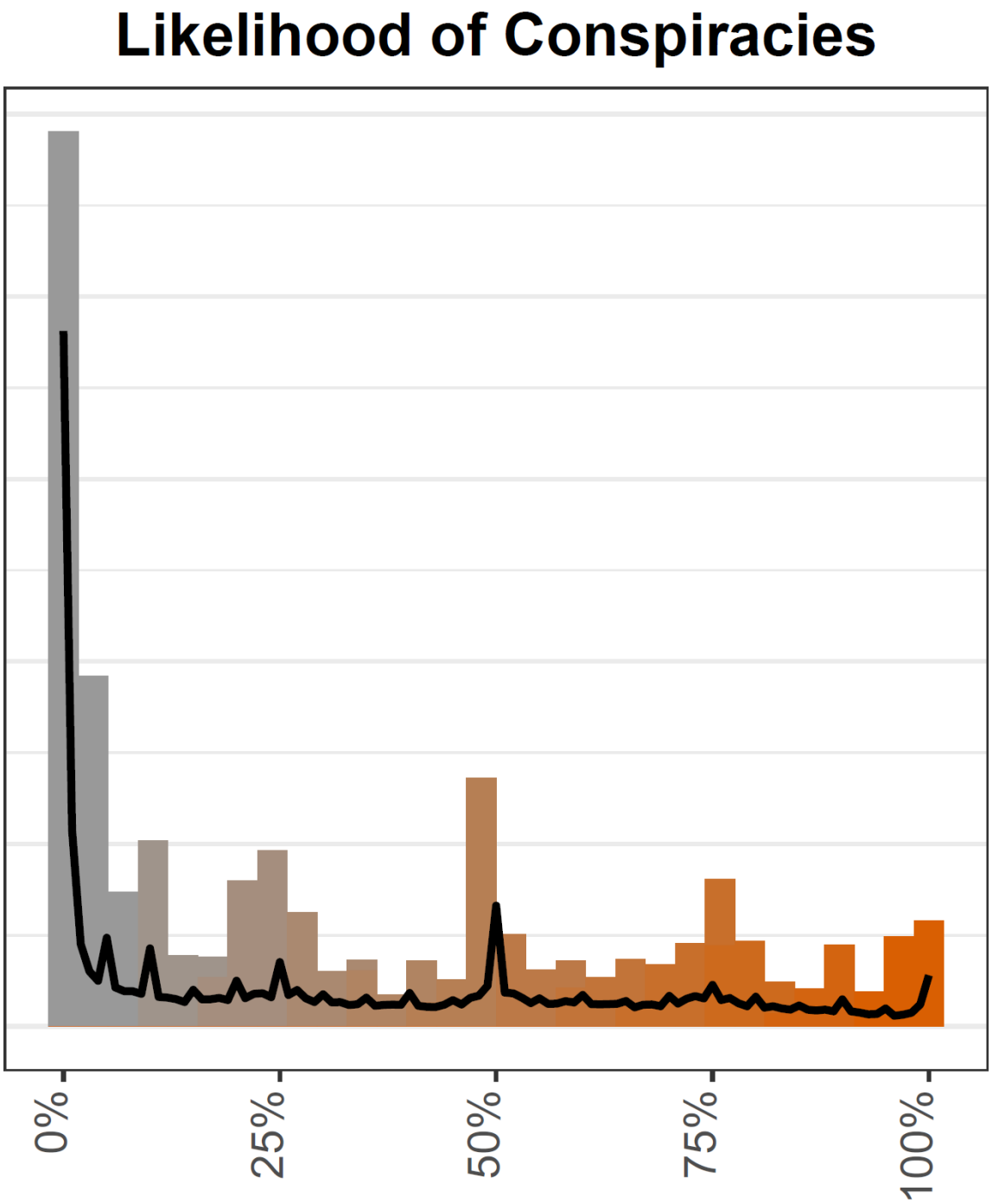

Figure 4-6: Weighted likelihood responses across all sentences binned by whether the sentences are non-evidence-based. The unweighted data is in the form of a line. The most common response by far is a total rejection of misinformed statements. When a misinformed statement is judged as possible, there is a large amount of variance in responses, indicating high uncertainty which implies these beliefs could be self-correcting over time. Very few responses indicate a non-evidence-based statement is 100% certain, signaling intellectual humility. Note that weighing the raw data raises prevalences relatively uniformly.

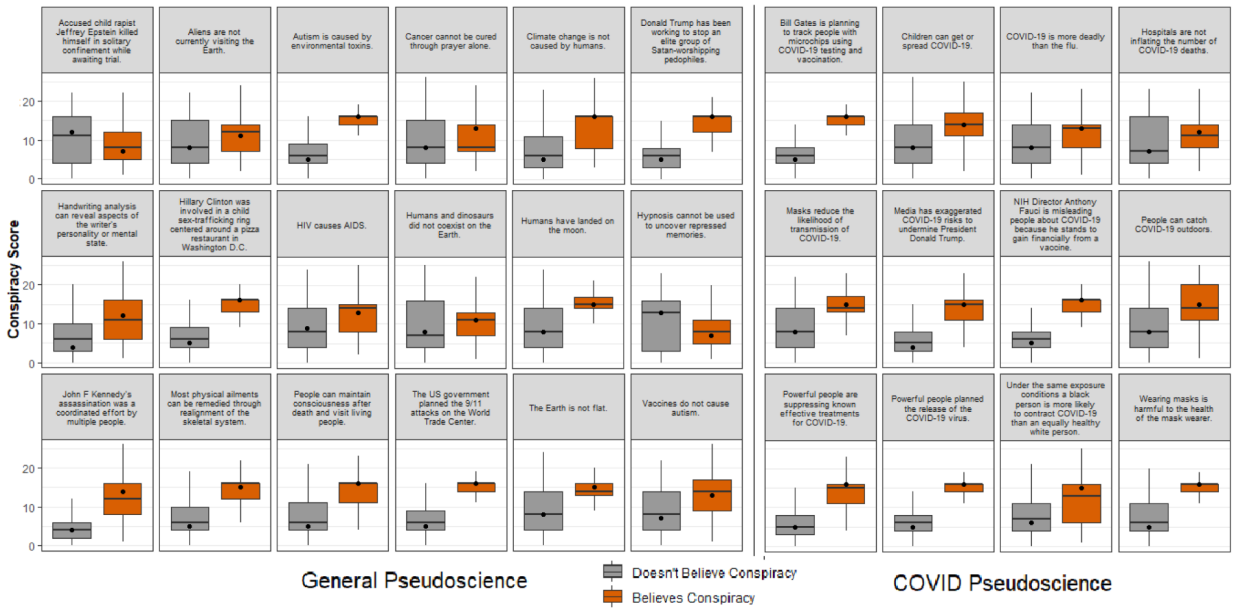

Figure 4-7: Conspiracy scores (total number of misinformed statements believed) binned by whether participants believe a particular statement. Boxplots are weighted data, black dots are the unweighted medians. Orange boxes represent individuals who believe a nonevidence-based statement while gray boxes represent non-believers. Boxes show the median 50% quantiles of conspiracy scores. Non-overlapping boxes within a statement indicate 95% confidence that the true medians differ. Most box pairs overlap indicating an overall lack of support for the slippery slope theory.

Discussion

Summary

Although any given misinformed belief may be uncommon, in aggregate these beliefs are extremely common. The median number of non-evidence-based beliefs held by individuals is 9, and 98% of people believe at least one. These beliefs are almost always accompanied by a degree of uncertainty however.

We find only modest evidence to support the slippery slope theory. Most endorsed beliefs do not show a significantly higher conspiracy score. Only when looking at closely related beliefs such as beliefs related to COVID-19, does belief in one statement tend to predict belief in the others, signaling that individuals do not view all theories equally. Uncovering more predictors of beliefs, such as demographics and higher-order beliefs, should be the focus of future research.

Implications

We have demonstrated that misinformed beliefs are widespread throughout the population, but often weakly held. This pattern is consistent with the idea that people are trying their best to understand what is true in the world but are making errors. Since uncertainty is associated with higher degrees of interest and curiosity, we would expect

these particular misinformed beliefs to be more likely to update and self-correct as further evidence is sampled and integrated. Thus, these kinds of misinformed beliefs—those which are held with higher degrees of uncertainty—are less problematic in context because we would not expect them to be stubbornly held.

Further evidence in support of this idea is our finding that misinformed COVID-19 beliefs were less common than misinformed beliefs at large. If this measurable difference can be directly linked to the efficacy of public policy education efforts, it will provide clear evidence that effective interventions to combat misinformation should provide people with clear, specific evidence. We suspect that, given the patterns we observed here, these approaches may be more effective than generally trying to train people against dubious belief adoption at large.

Social prevalence encourages misinformed beliefs

The “illusory truth effect" says that people will tend to believe something more strongly the more they encounter it (Hasher et al., 1977). The effect starts with as little as two exposures, has been found in children as young as 5 (Fazio & Sherry, 2020), and is just as strong regardless of an individual’s cognitive ability, need for cognitive closure, or cognitive style (De Keersmaecker et al., 2020). Compounding this effect is evidence that individuals overestimate the amount of information they use to form a belief (Klein & O’Brien, 2018). These facts—in tandem with their relationship to relatively sudden recent changes in how people gather their information from the world—may help explain the prevalence we observe here.

We note that the most commonly endorsed misinformed beliefs in our study are, by definition, not fringe beliefs. Certain statements, such as the belief that handwriting analysis can reveal an individual’s personality, are endorsed not only by certain authorities and popular culture, but by the majority of people in the world (70% in our sample). Since we know that social prevalence is one important cue people use in order to infer what is true (Orticio et al., 2021), the more people already believe this kind of non-evidence-based belief, the more we’d expect will believe it in the future. This isn’t irrational given the limitations of human access to truth. Since no one has direct access to truth, the best a person can do is engage in inference from sparse, often indirect sources of evidence, like the opinions of others.

Misinformed beliefs in informational ecosystems

The changing way in which people sample evidence from the world in order to infer truth is important to consider in light of our results. People increasingly rely on online sources for information, and almost all recommend content based on algorithms that maximize engagement (clicks, likes, comments, or hang time on the visual image). Maximizing engagement likely results in promoting material that is less likely to be true. For example, recent work suggests interestingness-if-true is a strong predictor of news sharing, not the user’s assessment of its likelihood of actually being true (Altay et al,. 2021). This is important because we know that estimates of social prevalence increase the likelihood of the adoption of misinformation as belief (Orticio et al., 2021).

The engagement-based reward systems used by platforms like YouTube, TikTok, and Twitter to incentivize content creation also likely add bias to the information pool. Creators are rewarded with views, likes, or money proportional to how many sets of eyes their content attracts and the duration they keep them glued—not the veracity of what they post. These pressures likely incentivize the creation of sensationalized, conspiratorial, and fringe content due to its novelty and subsequent interest.

If our information ecosystem is polluted, it is particularly problematic in light of human fallibility. People have built-in mechanisms designed to help them sparsely sample information in the world to draw quick inferences and act (Kidd & Hayden, 2015; Wade & Kidd, 2019). People seek out information to reduce their uncertainty, but move from uncertain to relatively sure on the basis of heuristics like feedback (Martí et al., 2018). Once certain, people tend to stick stubbornly with their established beliefs, and it is difficult to prompt them to revise. These cognitive mechanisms are useful in preventing people from wasting time from material that is already understood.

However, this aspect of human psychology may be problematic in environments with misleading feedback signals and which offer several points of related feedback quickly, as is true when people seek answers online.

Materials and Methods

We recruited 2,036 participants using a custom built web interface on Amazon Mechanical Turk on November 24, 2020. We required participants to be from the U.S. and have at least a 95% approval rating from previous tasks. Responses were recorded on a secure server. After consenting to the experiment, participants were asked to type out a series of sentences, pledging to answer questions honestly. This was followed by a nine question demographics questionnaire (age, sex, race, ethnicity, state of residence, education, income, religion, and politics). Next, participants entered two practice trials where they rated the likelihood of statements (“Plants need water to grow." and “Birds lay eggs.") using a slider bar, and also guessed how many other people would find them to be probably or definitely true. Then, participants saw 60 more trials of the same format (see Figure 4-1). In order to test for differences between a broad-range of misinformed beliefs and those within a more narrow scope, each of these trials displayed one of 18 general statements (half non-evidence-based, half factual) or 12 COVID-19-relevant statements (seven non evidence-based, five factual). Each statement was presented twice for a total of 60 trials as a way of assessing reliability. After every 15 trials, a free-response catch question was asked (“What is your favorite drink?", “What is your favorite movie?", “What is your favorite snack?", “What is your favorite aquatic animal?") to be used to filter careless or automated responses from polluting our participant pool.

Analysis

We used our data to estimate overall belief prevalence in the U.S. by correcting the sampling bias in our Amazon Mechanical Turk data. The Amazon Mechanical Turk’s participant pool tends to be more white, male, young, and poor than the general U.S. population (Moss et al., 2020).

To correct for this sampling bias, we employed an “iterative proportional fitting", or “raking" technique (Deming & Stephan, 1940). Raking applies a weight to each participant to offset sampling bias. For example, if the true proportion of males in the population is 50% but your sample is only 25% male, raking will apply a weight of 3 to each male and 1 to each female. If this process is performed for more than one variable, adjusting a participant’s weight to match the true proportions for one variable may ruin the weight value for another variable. To correct for this, the algorithm is run for many iterations and only stops when all weighted proportions are within a set threshold (ϵ) of their true proportions. Our algorithm ran for 100 iterations with ϵ = .000005.

The true proportions for age, sex, race, ethnicity, state of residence, education, and income were calculated using the 2014–2018 American Community Survey Public Use Microdata Sample (PUMS) from the United States Census Bureau, after excluding all individuals under 18. The true proportions for religious and political affiliations were taken from the Pew Research Center 2014 Religious Landscape Study. We followed raking best practices (Battaglia et al., 2009) including combining certain demographic categories with extremely few entries in our data. The combined categories are “American Native or Other" for race, “Non-Mexican Hispanic/Latino" for ethnicity, “Master’s/Doctorate/Professional degree" and “No High School Diploma or GED" for education, “Mainline Protestant", “Other Non-Christian", “Other Christian", and “Nothing in particular (religion not important)" for religion, and “Moderate" for politics.

Exclusion criteria

We applied a conservative exclusion criteria to our data which maintained 1,717 participants and excluded those who demonstrated inattention or who failed to demonstrate their humanity. We excluded 195 participants for not following the pledge-typing instructions at the onset of the experiment and 124 participants for not providing at least 3 out of 4 valid catch question responses.

For the remaining 1,717 participants, we examined their reliability by first labeling a repeated sentence as a bad trial if the participant gave likelihood scores that were greater than or equal to 20 points apart (out of 100). Using this criteria, the mean reliability for our participants is 88.7%, meaning 88.7% of all trial pairs were rated as less than 20 points apart. Note that even with a 20-point criteria (which is very conservative) our observed reliability remains very high. We do not exclude any participants with low reliability since it is possible that different likelihood scores on the same items is a reflection about their uncertainty, not about a lack of attention.[8]

Acknowledgements

We would like to thank members of the Kidd Lab for providing valuable feedback, especially Sarah Stolp, David O'Shaughnessy, and Holly Palmeri.

Funding

[8] Applying reasonable reliability exclusion criteria ends in very few participants being excluded and does not change our overall results.

DARPA (Machine Common Sense TA1, BAA number HR001119S0005, CK)
Berkeley Center for New Media (CK)
The Jacobs Foundation (CK)
Georgia Lee Fellowship through the Hellman Scholars Fund (CK)

Author contributions

Conceptualization: LM, AC, CK
Methodology: LM, CK
Investigation: LM
Software: LM
Formal analysis: LM
Visualization: LM, CK
Writing – original draft preparation: LM
Writing – review and editing: LM, AC, CK
Supervision: CK
Funding acquisition: CK

V. Discussion

The consequences of misinformed beliefs

The objective of my line of research is to understand both the reasons we become certain and the subsequent beliefs that we hold as a consequence. Beliefs form the basis of our decisions and actions, and thus misinformed beliefs can have serious consequences when they lead to unjustified actions.

On December 4, 2016, a 28-year-old man from North Carolina arrived at a pizza restaurant in Washington D.C. and fired three gunshots which struck the restaurant's walls, a desk, and a door. He was certain that the restaurant was holding child sex slaves in their basement and wanted to confirm for himself they were there. He believed he was going to be a hero who was about to rescue children from captivity. There were no such children, however, and he was soon arrested by police. The shooter had previously seen online materials promoting a conspiracy theory which claimed the Democratic party was linked to a pedophilia ring that involved Satanic ritual abuse and held meetings at the pizza restaurant. Viewing content from internet message boards and social media left the gunman with a sense of certainty that was sufficiently high enough that he loaded a rifle and drove 360 miles to the pizza restaurant.

Humans' metacognitive abilities to reflect on the accuracy of beliefs is central to efficiently building accurate models of the world. When certainty is appropriately high, it can give us assurances we are pursuing a correct course of action. When it is appropriately low, it can signal we need to change course. Yet, this dissertation contains empirical data detailing cases where certainty is miscalibrated to reality. If certainty is inappropriately low, we may miss out on learning valuable new information. If certainty is inappropriately high, we run the risk of acting on potentially dangerous misinformation.

Feedback primarily determines certainty

Chapter 2 presented experiments in which 38 behavioral and model-based predictors were tested in order to uncover which predicted certainty the best. Across four experiments in a high-level concept learning task, participants consistently used behavioral predictors over model predictors, indicating certainty was being derived from a fast, but error-prone, heuristic over a more costly but accurate model-based calculation. More specifically, the best predictor was the participant's accuracy in the last several trials. While this might be a decent guidepost as to whether or not you should be certain in the aggregate, given that chance in the task was 50%, a participant who possessed an incorrect concept, but happened to answer correctly over several trials, would be highly certain about a falsehood. Similarly, if they possessed the correct concept, but made errors and answered incorrectly, their certainty would be much lower than it should be. Importantly, these results show that although humans are primarily using heuristics, model-based predictors are making unique contributions to certainty. Thus, individuals are using many different lines of evidence when forming certainty, not

simply one. Overall these findings point to a complex, but flawed, process of certainty formation which is likely a major source of inaccurate beliefs.

While certainty in low-level perceptual domains has been found to be well calibrated to reality (Barthelmé & Mamassian, 2009; Sanders et al., 2016), the findings detailed in Chapter 2 point to a miscalibration in high-level domains. Given this domain sensitivity, it is likely that Chapter 2's specific pattern of results only applies to high-level domains. Future work should examine these predictors in a variety of low-level and high-level domains in order to better understand how certainty is formed across different domains. Modifying low-level and high-level tasks might also result in shifts to behavioral/model predictor preferences. In other words, it may be possible to rely primarily on heuristics in a low-level task or on model-based predictors in a high-level task by modifying task features. These findings could result in interventions in order to make our sense of certainty more reliable. It is also unknown whether these findings are developmentally stable, or whether children will form their certainty through different means. A developmentally interesting finding would be if children were found to use model-based predictors over behavioral predictors. This would suggest that more objective predictors are innate while heuristics are learned strategies.

Humans underestimate conceptual diversity

Shared lexical concepts are a fundamental necessity for communication between people. These concepts facilitate the establishment of common ground and provide shortcuts when exchanging information. The common assumption is that except in cases of novel concepts, humans possess shared concepts. Two experiments detailed in Chapter 3 provide estimates that around a dozen distinct concepts exist in the population for common animals and famous politicians. Convergent estimates were obtained via two different paradigms, a similarity task, and a feature classification task. When asked to guess the number of people that agreed with their judgements, individuals tended to display a strong egocentric bias, vastly overestimating the number of people that agreed with them. Together, these findings suggest that concepts across the population are diverse, and that people are largely unaware of this diversity. This has implications for the way we communicate, teach, and learn. If individuals possess sufficiently different concepts, but believe they share common-ground, miscommunication may occur. This may not be a large issue when discussing what to eat for lunch, but may have larger implications when legislating laws which affect millions of individuals.

Future work should examine the reasons and ways in which concepts differ. Chapter 3's experiments only tested English speakers, and therefore, the amount of conceptual diversity is still unknown in other languages. The reasons for variations within a language should also be tested, with particular attention paid to regional, ethnic, socio-economic, and educational differences. Lastly, although there is evidence that these differences propagate upward to the level of societal disagreements (e.g. equity, abortion, justice), future studies should examine precisely how conceptual differences result in miscommunication and disagreement.

Misinformed beliefs are widely held

Past research into misinformed beliefs have focused on individual differences that correlate with conspiratorial thinking. When seeking out new information, dogmatic individuals tend to search less than non-dogmatic individuals (Schultz et al., 2020). Agreeableness and conscientiousness both appear to be modestly negatively correlated with belief in conspiracies, while distress, immodesty, impulsivity, and negative affect are all positively correlated (Bowes et al., 2021). Conservatives tend to not only believe in specific conspiracies more often than liberals, but also tend to endorse conspiratorial worldviews more often (van der Linden et al., 2021). Lower cognitive ability has also been found to be positively linked with false memories about a fake news event (Murphy et al., 2019).

These studies have mostly focused on a relatively narrow set of conspiratorial beliefs. In contrast, the evidence presented in Chapter 4 not only assessed 30 vastly different misinformed beliefs, but also assessed people's strength of certainty in those beliefs. The finding that the median individual believes 9 out of 30 misinformed beliefs demonstrates that there is not a subpopulation of particularly gullible conspiracy enthusiasts, but rather that the unusual subpopulation are the evidence-based skeptics who believe 0 out of 30 and number 47 individuals out of our sample of 1,717. Although these beliefs are pervasive, they are also often weakly held, with only a small percentage being rated as 100% likely, and most rated at 50% likely or less. Thus, although these beliefs are widely held, individuals overwhelmingly possess some degree of uncertainty about them. These findings have implications for science education and public policy. It may not be possible to "eradicate" a particular misinformed belief by developing a vaccine, but lowering certainty may be the next best thing.

Future research should focus on the complexity of the beliefs cataloged in Chapter 4. Examining higher-order beliefs could better uncover relationships between misinformed beliefs and lead to a better understanding of how these beliefs form. An individual who holds the general belief that governments hide the truth from the public might be more likely to believe certain beliefs such as 9/11 was an inside job and that Epstein was murdered. Similarly, investigating demographic predictors for classes of misinformed beliefs could help public policy efforts to minimize the spread of harmful beliefs by providing evidence on what demographics to focus on. Lastly, additional work examining the relationship between the certainty of a belief and efforts to "debunk" a misinformed belief would help understand why certain individuals seem particularly resistant to belief revision.

Implications

My work investigated people's sense of certainty in multiple high-level domains to uncover how certainty is derived and how it can lead to misinformed beliefs. This was done via behavioral experimentation and computational modeling for abstract concept learning, word meanings, and high-level theories about the world. This thesis presented evidence that adults tend to use simplified heuristics over more accurate calculations when calculating their certainty, which likely leads to unjustifiably high confidence in particular situations in which feedback is concentrated.

Misinformed beliefs can lead to hurtful and often dangerous behaviors. These beliefs are likely formed from exposure to information ecosystems which result in an unearned sense of certainty. Forms of concentrated, repeated feedback such as media outlets which broadcast that the Democratic party is holding paedophilic satanic rituals in the basement of a pizza restaurant will inevitably lead listeners to high amounts of certainty about the veracity of the claims. Because it is more difficult to shift beliefs once they are firmly held, it becomes far more important that we prevent misinformed beliefs from taking hold in the first place. With the proliferation of internet access and the ability to “do your own research”, access to unreliable information sources with ulterior motives is more impactful than ever. Epistemically naive individuals can be led to science-based knowledge or they can be led to Pizzagate.

References

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: critique and reformulation. *Journal of Abnormal Psychology, 87*(1), 49.

Altay, S., de Araujo, E., & Mercier, H. (2021). “If this account is true, it is most enormously wonderful”: Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, 1-22.

Amarasingam, A., & Argentino, M. A. (2020). The QAnon conspiracy theory: A security threat in the making. *CTC Sentinel*, *13*(7), 37-44.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making, 2*(2), 81-94.

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281-311.

Baron, J. (2008). Thinking and Deciding, 4th edition. *New York: Cambridge University Press.*

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*, 255–278.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge University Press.

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

Battaglia, M. P., Hoaglin, D. C., & Frankel, M. R. (2009). Practical considerations in raking survey data. *Survey Practice*, *2*(5), 2953.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new jim code. *Social Forces*.

Bocci, L., & Vichi, M. (2011). The K-INDSCAL model for heterogeneous three-way dissimilarity data. *Psychometrika*, *76*(4), 691-714.

Bowes, S. M., Costello, T. H., Ma, W., & Lilienfeld, S. O. (2021). Looking under the tinfoil hat: Clarifying the personological and psychopathological correlates of conspiracy beliefs. *Journal of Personality, 89*(3), 422-436.

Bruner, J. S., & Austin, G. A. (1986). A study of thinking. Piscataway, NJ: Transaction.

Bush, L. E. (1973). Individual differences multidimensional scaling of adjectives denoting feelings. *Journal of personality and social psychology*, *25*(1), 50.

Callaway, E. (1983). The pharmacology of human information processing. *Psychophysiology, 20*(4), 359-370.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. *Psychometrika*, *35*(3), 283-319.

Casasanto, D., & Lupyan, G. The conceptual mind: New directions in the study of concepts pp. 543–566 (2015).

Chao, A., & Chiu, C. H. (2016). Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, *1*, 26.

Clark, H. H. (1998). Communal lexicons. *Context in language learning and language understanding*, *6387*.

Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *345*(1311), 101-118.

Daniel, D. B., & Klaczynski, P. A. (2006). Developmental and individual differences in conditional reasoning: Effects of logic instructions and alternative antecedents. *Child Development, 77*(2), 339-354.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience, 16*(1), 105-110.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427-444.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*(1), 28-38.

De Keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2020). Investigating the robustness of the illusory truth

effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin, 46*(2), 204-215.

De Waal, E., & Ten Hagen, S. L. (2020). The Concept of Fact in German Physics around 1900: A Comparison between Mach and Einstein. *Physics in Perspective*, *22*.

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*, 403–408.

Evans, J. S. B. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning, 13*(4), 321-339.

Evans, J. S. B., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382-389.

Fazio, L. K., & Sherry, C. L. (2020). The effect of repetition on truth judgments across development. *Psychological Science*, *31*(9), 1150-1160.

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

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911.

Frimer, J. A. (2020). Do liberals and conservatives use different moral languages? Two replications and six extensions of Graham, Haidt, and Nosek's (2009) moral text analysis. *Journal of Research in Personality*, *84*, 103906.

Gale, W. A., & Sampson, G. (1995). Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*(3), 217-237.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). *Chapman and Hall/CRC*.

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual Review of Anthropology*, *40*, 379-398.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1-12.

Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives, 7*(3), 160-165.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences, 7*(1), 43-48.

Goldfarb, N. (2021). The Use of Corpus Linguistics in Legal Interpretation. *Annual Review of Linguistics*, *7*, 473-491.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press on Demand.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3-4), 237-264.

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

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 75-86.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Grossnickle, E. M. (2016). Disentangling curiosity: Dimensionality, definitions, and distinctions from interest in educational contexts. *Educational Psychology Review*, *28*(1), 23-60.

Hampton, J. A., & Passanisi, A. (2016). When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(4), 505.

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In Psychology of learning and motivation (Vol. 62, pp. 33-58). Academic Press.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal behavior, 16*(1), 107-112.

Hernandez, I., & Preston, J. L. (2013). Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology, 49*(1), 178-182.

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

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

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

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449-460.

Klein, N., & O'Brien, E. (2018). People use less information than they think to make up their minds. *Proceedings of the National Academy of Sciences*, *115*(52), 13222-13227.

Koriat, A., & Sorka, H. (2015). The construction of categorization judgments: Using subjective confidence and response latency to test a distributed model. *Cognition, 134,* 21-38.

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,* 1121–1134.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658-676.

Labov, W. (1973). The boundaries of words and their meanings. In C. Bailey & R. Shuy (Eds.), *New ways of analyzing variation in English* (pp. 340-395). Georgetown University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Landrum, A. R., Olshansky, A., & Richards, O. (2021). Differential susceptibility to misleading flat earth arguments on youtube. *Media Psychology*, 24(1), 136-165.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2), 311.

Langer, E. J., & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology, 32*(6), 951.

Lefcourt, H. M. (1973). The function of the illusions of control and freedom. *American Psychologist, 28*(5), 417.

Lehrer, K. (2018). *Theory of knowledge*. Routledge.

Leshem, O. A., & Halperin, E. (2020). Lay theories of peace and their influence on policy preference during violent conflict. *Proceedings of the National Academy of Sciences*, *117*(31), 18378-18384.

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*, 607–616.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170.

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1.

Margolis, E., & Laurence, S. (Eds.). (1999). *Concepts: Core readings.* The MIT Press.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431-467.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological bulletin*, *102*(1), 72.

Martí, L., Mollica, F., Piantadosi, S., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind, 2*(2), 47-60.

Martinez-Berman, L., McCutcheon, L., & Huynh, H. P. (2021). Is the worship of celebrities associated with resistance to vaccinations? Relationships between celebrity admiration, anti-vaccination attitudes, and beliefs in conspiracy. *Psychology, Health & Medicine, 26*(9), 1063-1072.

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology, 102*(3), 392-405.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*(4), 462-472.

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,* 508–520.

McGee, V. E. (1968). Multidimensional scaling of N sets of similarity measures: A nonmetric individual differences approach. *Multivariate Behavioral Research*, *3*(2), 233-248.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Medin, D. L., & Rips, L. J. (2005). Concepts and Categories: Memory, Meaning, and Metaphysics. *Cambridge University Press*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moss, A. J., Rosenzweig, C., Robinson, J., & Litman, L. (2020). Demographic stability on Mechanical Turk despite COVID-19. *Trends in cognitive sciences*, *24*(9), 678-680.

Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological Science, 30*(10), 1449-1459.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour, 1*(11), 810–818.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, *70*, 35-44.

Orticio, E., Martí, L., & Kidd, C. (2021). Beliefs are most swayed by social prevalence under uncertainty. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision Making, 15*(4), 476.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185-200.

Peterson, W. W. T. G., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, 4*(4), 171-212.

Pew Research Center 2014 Religious Landscape Study Pew Research Center, Washington, D.C. (2014).

Piantadosi, S. T. (2015). Problems in philosophy of mathematics: A view from cognitive science. In *Mathematics, Substance and Surmise* (pp. 305-320). Springer, Cham.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, *102*(2), 145-158.

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

Pretz, J. E., & Sternberg, R. J. (2005). Unifying the field: Cognition and intelligence.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369-381.

Rosch, E., & Lloyd, B. B. (Eds.). (1978). Cognition and categorization.
Sanchez, C., & Dunning, D. (2021). Jumping to conclusions: Implications for reasoning errors, false belief, knowledge corruption, and impeded learning. *Journal of Personality and Social Psychology*, 120(3), 789.

Romer, D., & Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social science & medicine*, *263*, 113356.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1-10.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of personality and social psychology*, *37*(3), 322.

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron, 90*(3), 499-506.

Schnuerch, M., Nadarevic, L., & Rouder, J. N. (2021). The truth revisited: Bayesian analysis of individual differences in the truth effect. *Psychonomic Bulletin & Review, 28*(3), 750-765.

Schulz, L., Rollwage, M., Dolan, R. J., & Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences*, *117*(49), 31527-31534.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125-140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390-398.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs: *General and Applied, 75*(13), 1–42.

Shtulman, A., Share, I., Silber-Marker, R., & Landrum, A. R. (2020). OMG GMO! Parent-child conversations about genetically modified foods. *Cognitive Development*, *55*, 100895.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3.

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*, 717–741.

Stanovich, K. E. (2004). Balance in psychological research: The dual process perspective. *Behavioral and Brain Sciences, 27*(3), 357-358.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-665.

Starmans, C., & Friedman, O. (2020). Expert or esoteric? Philosophers attribute knowledge differently than all other academics. *Cognitive Science*, *44*(7), e12850.

Jacobellis v. Ohio, 378 U.S. 184 (1964).

Stotz, K., Griffiths, P. E., & Knight, R. (2004). How biologists conceptualize genes: an empirical study. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 35(4), 647-673.

Sumner, E., DeAngelis, E., Hyatt, M., Goodman, N., & Kidd, C. (2019). Cake or broccoli? Recency biases children’s verbal responses. *PloS one*, *14*(6), e0217207.

Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, *42*(1), 7-67.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind.* Basic Books/Hachette Book Group.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401-419.

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.

Torres, M. N., Barberia, I., & Rodríguez-Ferreiro, J. (2020). Causal illusion as a cognitive basis of pseudoscientific beliefs. *British Journal of Psychology, 111*(4), 840-852.

Trifonov, E. N. (2011). Vocabulary of definitions of life suggests a definition. *Journal of Biomolecular Structure and Dynamics*, *29*(2), 259-266.

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.

Tversky, A. & Gati, I. (1978). Studies of similarity. In Eleanor Rosch & Barbara Lloyd (eds.), *Cognition and Categorization*. Lawrence Elbaum Associates.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185(*4157), 1124-1131.

Tucker, L. R., & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika*, *28*(4), 333-367.

U.S. Census Bureau., 2014-2018 american community survey 5-year public use microdata samples (2020).

Uscinski, J. E., Douglas, K., & Lewandowsky, S. (2017). Climate change conspiracy theories. In *Oxford Research Encyclopedia of Climate Science*.

van der Linden, S., Panagopoulos, C., Azevedo, F., & Jost, J. T. (2021). The paranoid style in American politics revisited: An ideological asymmetry in conspiratorial thinking. *Political Psychology, 42*(1), 23-51.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PloS One, 8*(5), e63507.

Wade, S., & Kidd, C. (2019). The role of prior knowledge and curiosity in learning. *Psychonomic Bulletin & Review, 26*(4), 1377-1387.

Wason, P. C., & Evans, J. S. B. (1974). Dual processes in reasoning?. *Cognition, 3(*2), 141-154.

Wigner, E. P. (1995). Review of the Quantum-Mechanical Measurement Problem. In *Philosophical Reflections and Syntheses* (pp. 225-244). Springer, Berlin, Heidelberg.

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics* (pp. 607-632). Oxford, England: Blackwell.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Winkler, A. (2011). *Gunfight: The battle over the right to bear arms in America*. WW Norton & Company.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*, 515–526.

Woko, C., Siegel, L., & Hornik, R. (2020). An investigation of low COVID-19 vaccination intentions among Black Americans: The role of behavioral beliefs and trust in COVID-19 information sources. *Journal of Health Communication*, *25*(10), 819-826.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, *3*(6), 767-773.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245.

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,* 558–563.

Young, O. (2019). What Role has Social Media Played in Violence Perpetrated by Incels?.

Zachos, F. E. (2016). Species concepts in biology (Vol. 801). Cham, Switzerland: Springer.

Appendices

Chapter II (Certainty Is Primarily Determined by Past Performance During Concept Learning) Appendices

Appendix II-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 2-1), 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[9]. 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 2-1), 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 2-1) 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 2-7).[10] 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 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.

---

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

[10] In order to test additional model-based predictors we also ran our model using a simplified grammar. See Table 2-8.

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[11] (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.

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

Appendix II-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 labeled with a “yes” or “no” to indicate whether it was part of the concept (see Figure 2-5). 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 2-6). Like Experiment 1, our model incorporated noise (alpha) and memory decay (beta) parameters, best fit as 0.65 and 0.06 respectively.

Appendix II-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 24 trials. Each condition tested for a different concept with varying complexity (see Table 2-1).

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.

Appendix II-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 labeled as “Not at all certain” and 5 labeled 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 24 trials. Each condition tested for a different concept with varying complexity (see Table 2-1).

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 2-14 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 2-15) 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 2-16).

Figure 2-17 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 2-13 shows the full model results for Experiment 4, sorted by AIC and giving the performance of each model in predicting certainty ratings.[12] 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

[12] See Table 2-14 for simplified grammar predictors

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.

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

Table 2-4: Predictors of certainty rankings for DNF grammar in Experiment 1 when analyzing data by participant vs. as a group. (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 2-5. 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
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

Table 2-6: Predictors of certainty for Experiment 1 using simplified grammar (behavioral predictors in gray).

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)

Table 2-7. Disjunctive normal form grammar used to generate logical rules in the idealized learning model. The variable x is the current object.

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 2-8: Simplified grammar

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

Table 2-9. Predictors of certainty for Experiment 2 (behavioral predictors in gray).

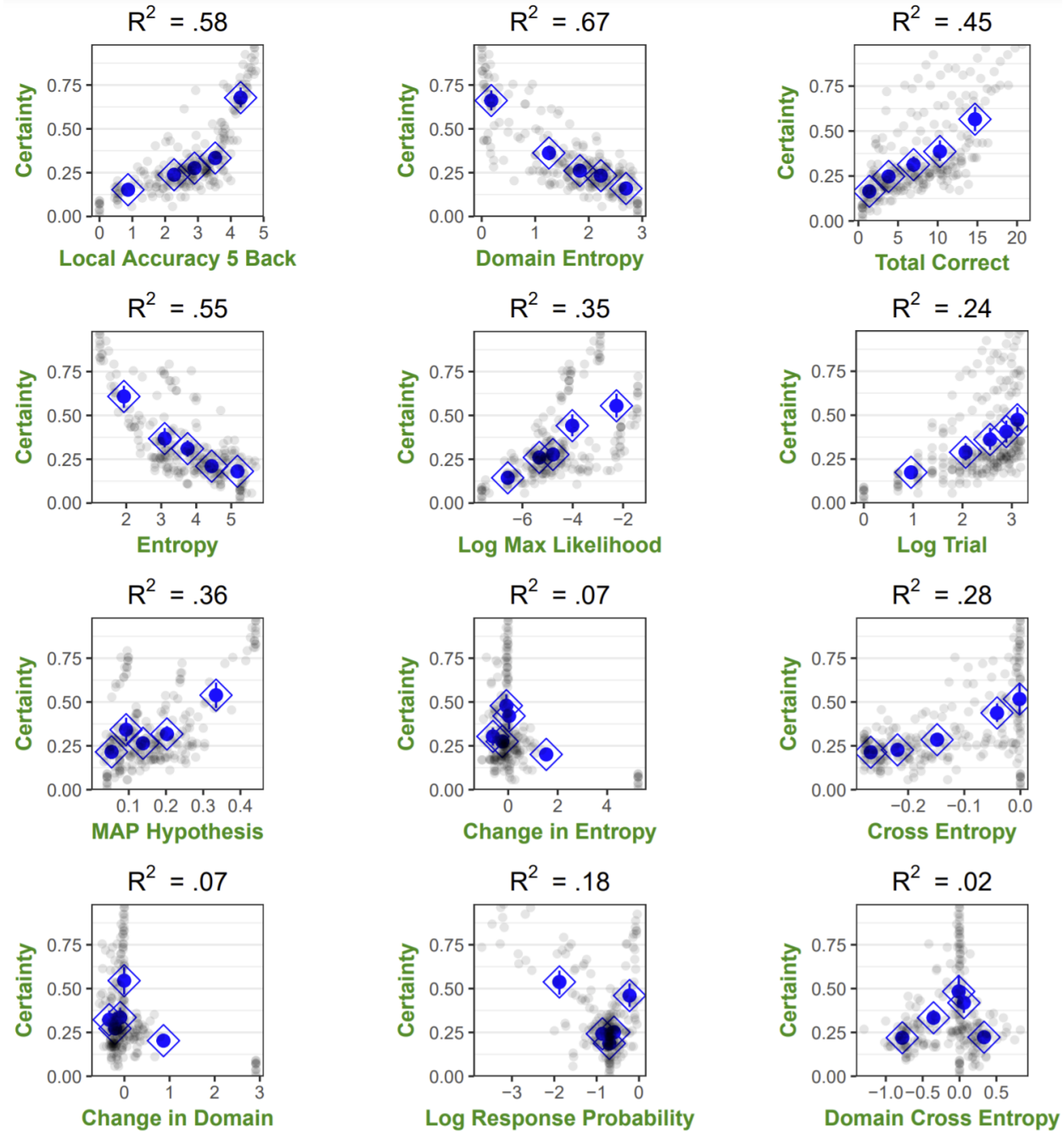

Figure 2-7: Key model fits for Experiment 1.

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	$< .001$

Table 2-10: Predictors of certainty for Experiment 2 using simplified grammar (behavioral predictors in gray).

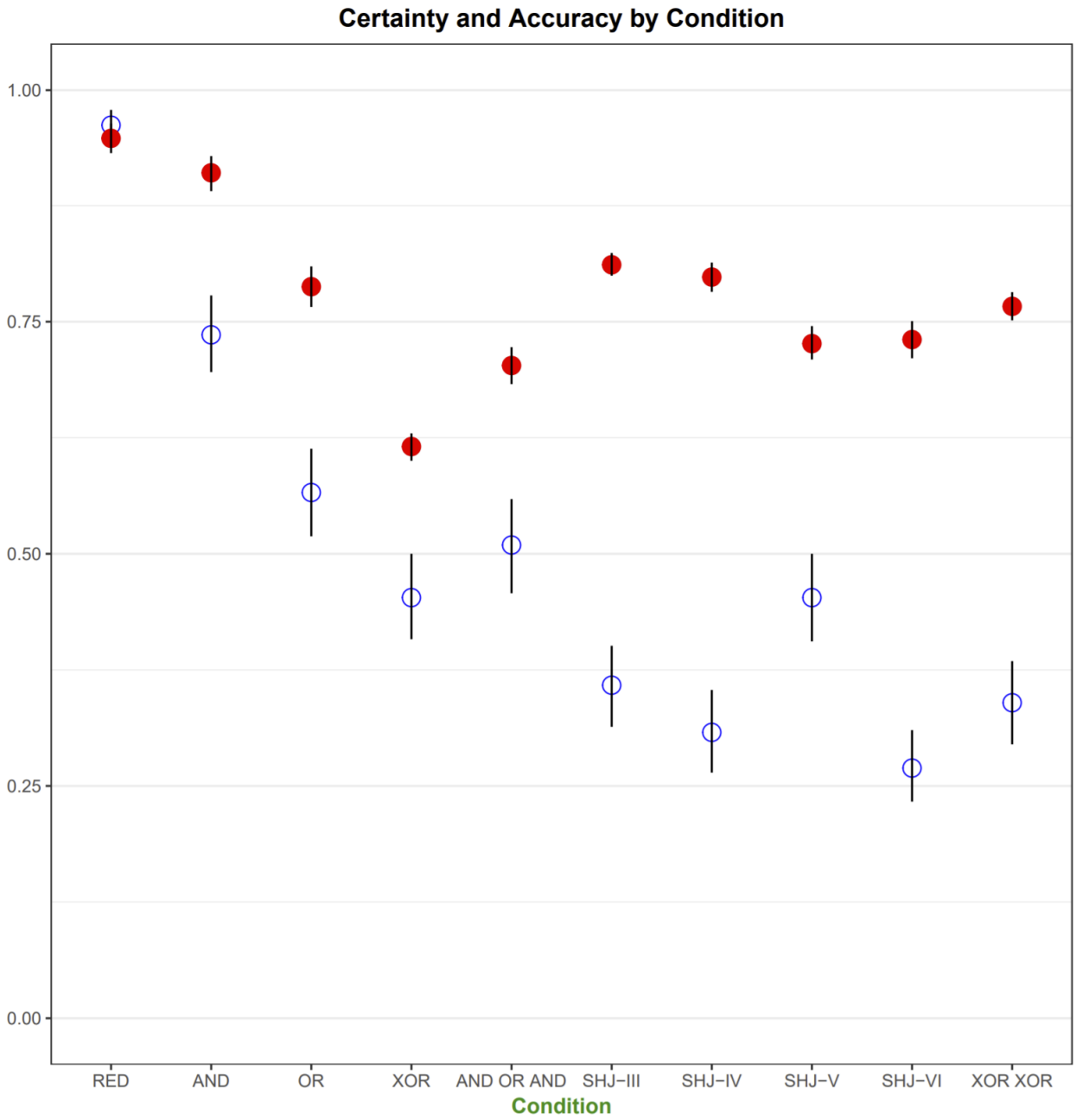

Figure 2-8. Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 2. Chance is 50% across all conditions if guesses are made randomly.

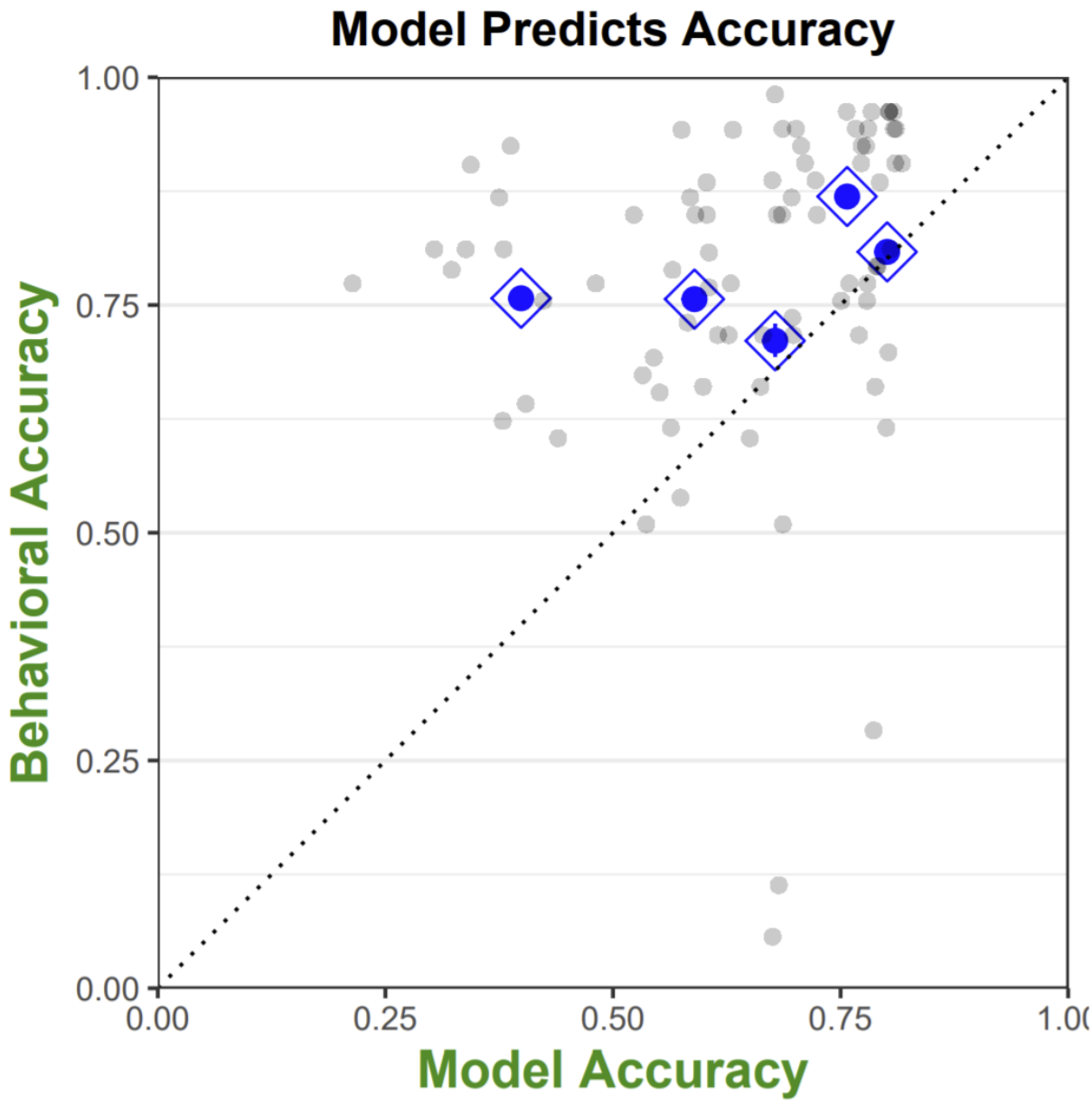

Figure 2-9: Model vs. behavioral accuracy for Experiment 2.

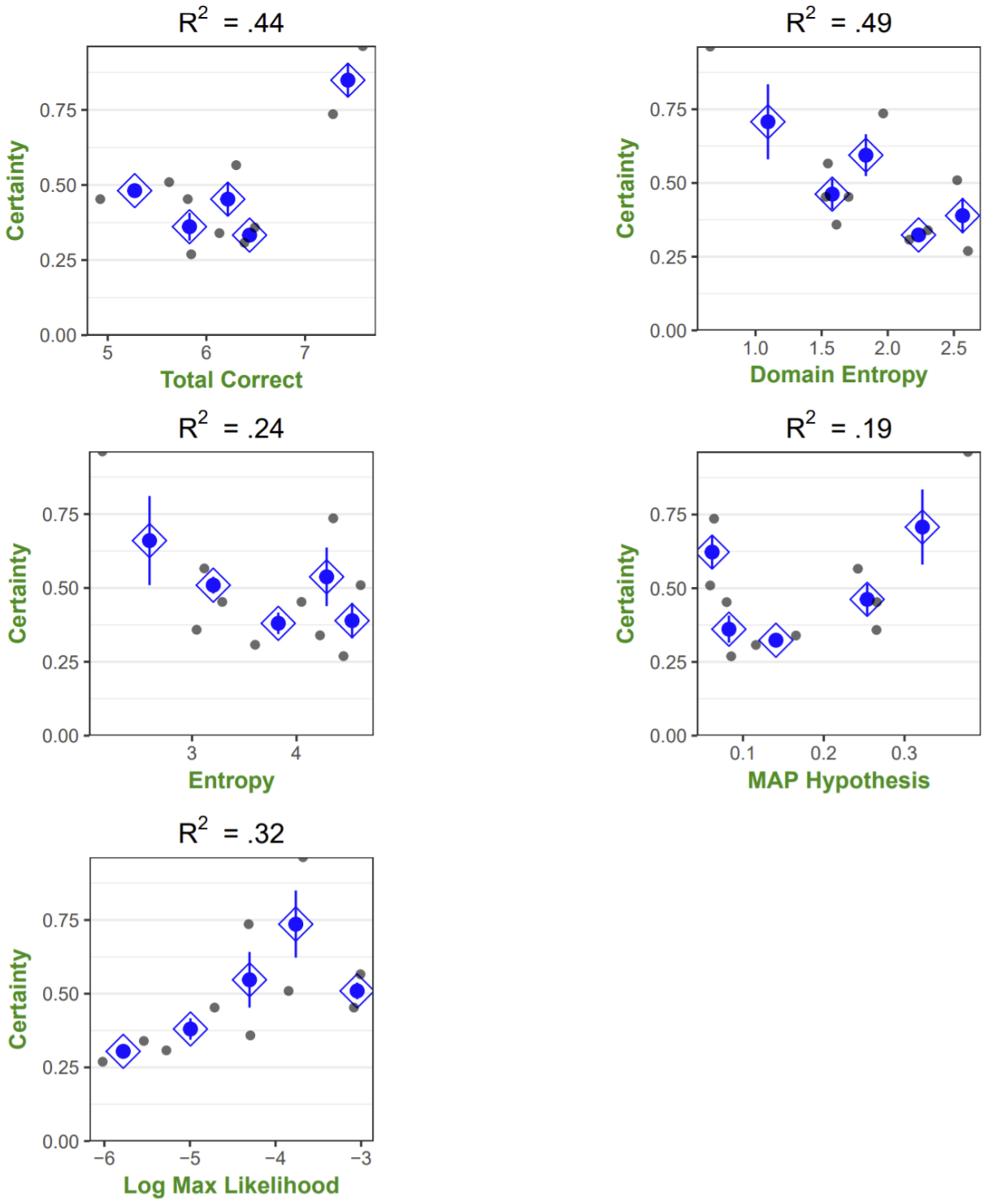

Figure 2-10. Key model fits for Experiment 2.

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

Table 2-11. 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
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

Table 2-12: Predictors of certainty for Experiment 3 using simplified grammar (behavioral predictors in gray)

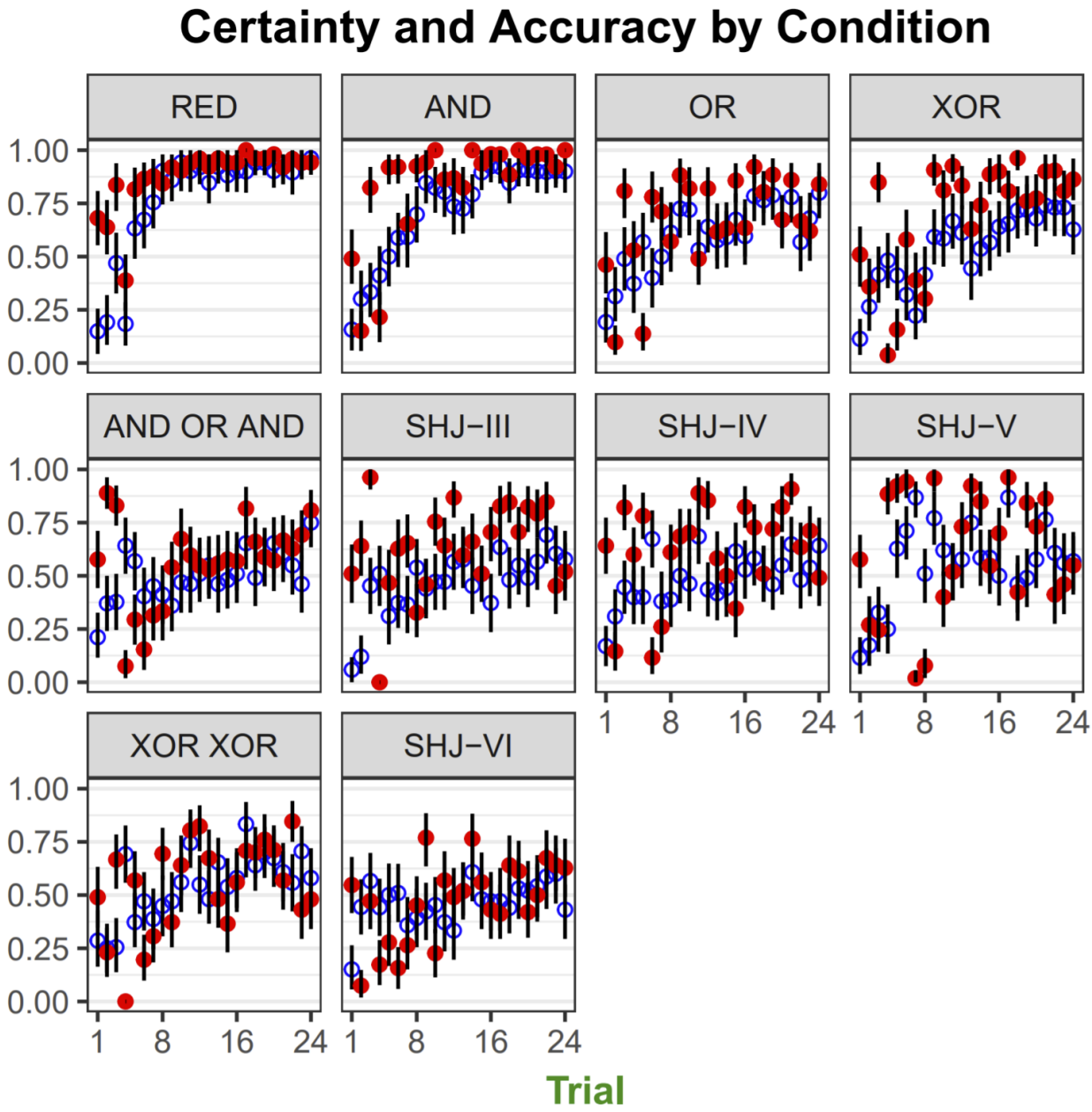

Figure 2-11: Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 3. Chance is 50% across all conditions if guesses are made randomly.

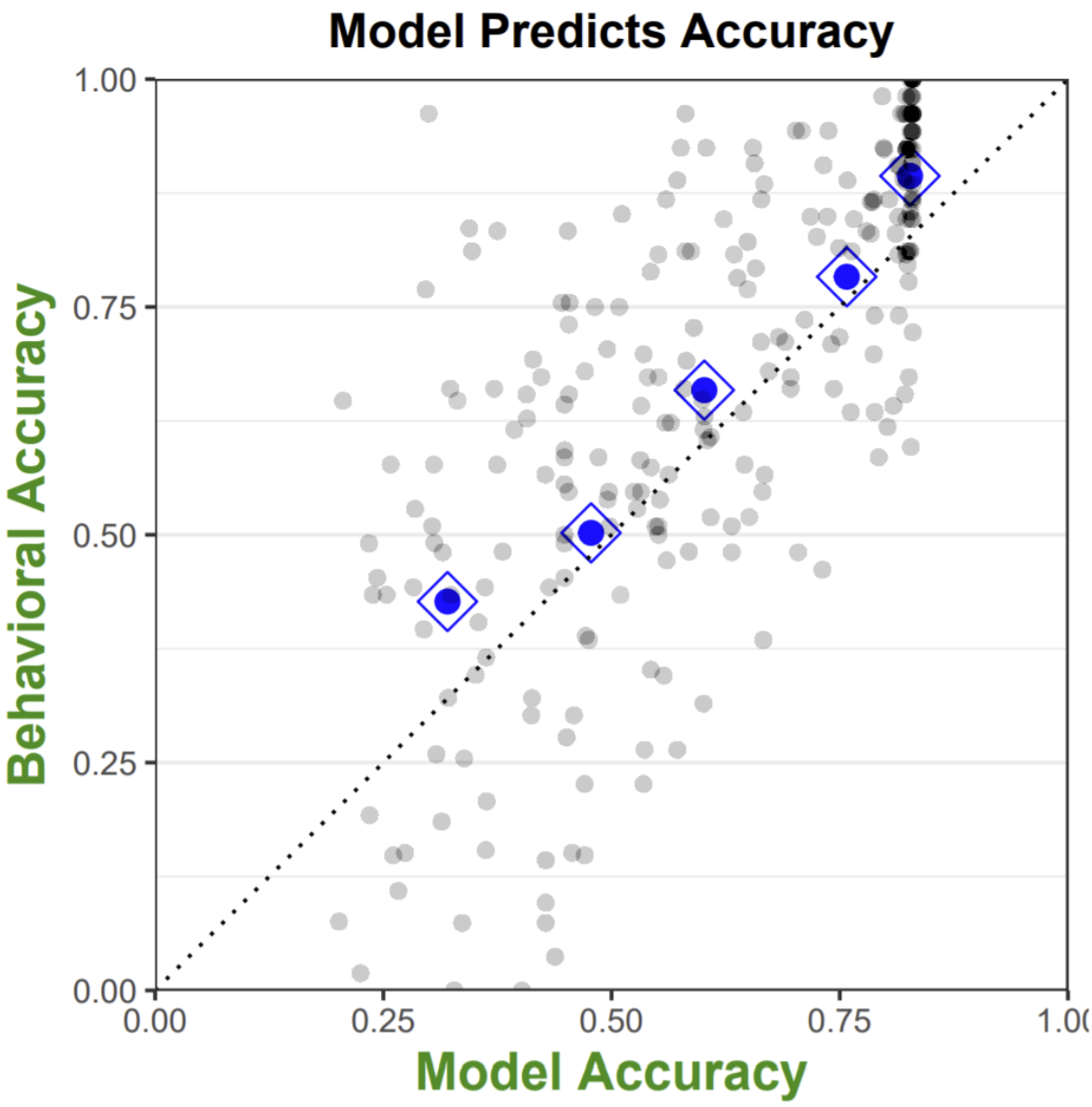

Figure 2-12. Model vs. behavioral accuracy for Experiment 3.

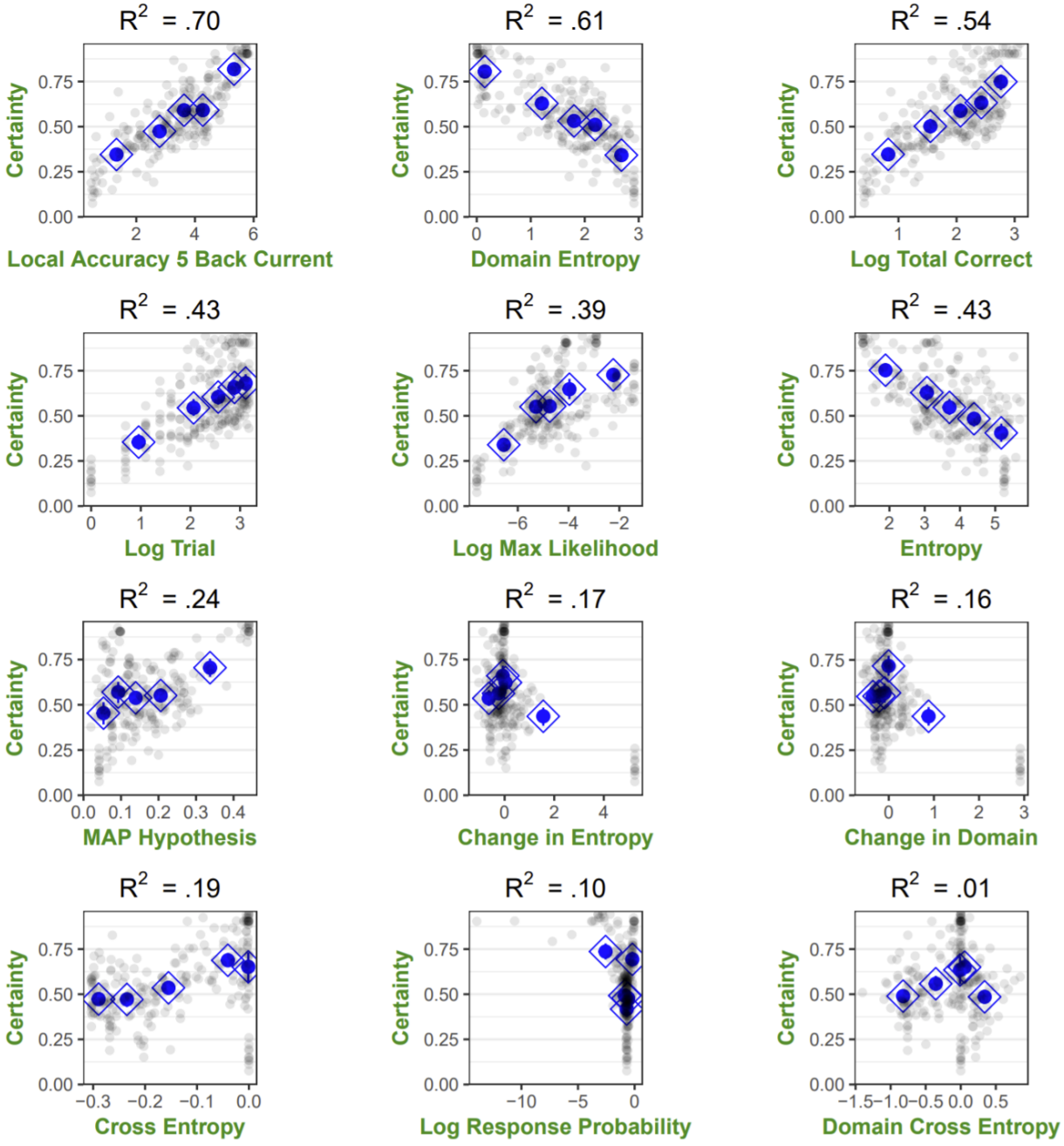

Figure 2-13: Key model fits for Experiment 3.

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

Table 2-13. 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	$< .001$
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	$< .001$

Table 2-14: Predictors of certainty for Experiment 4 using simplified grammar (behavioral predictors in gray).

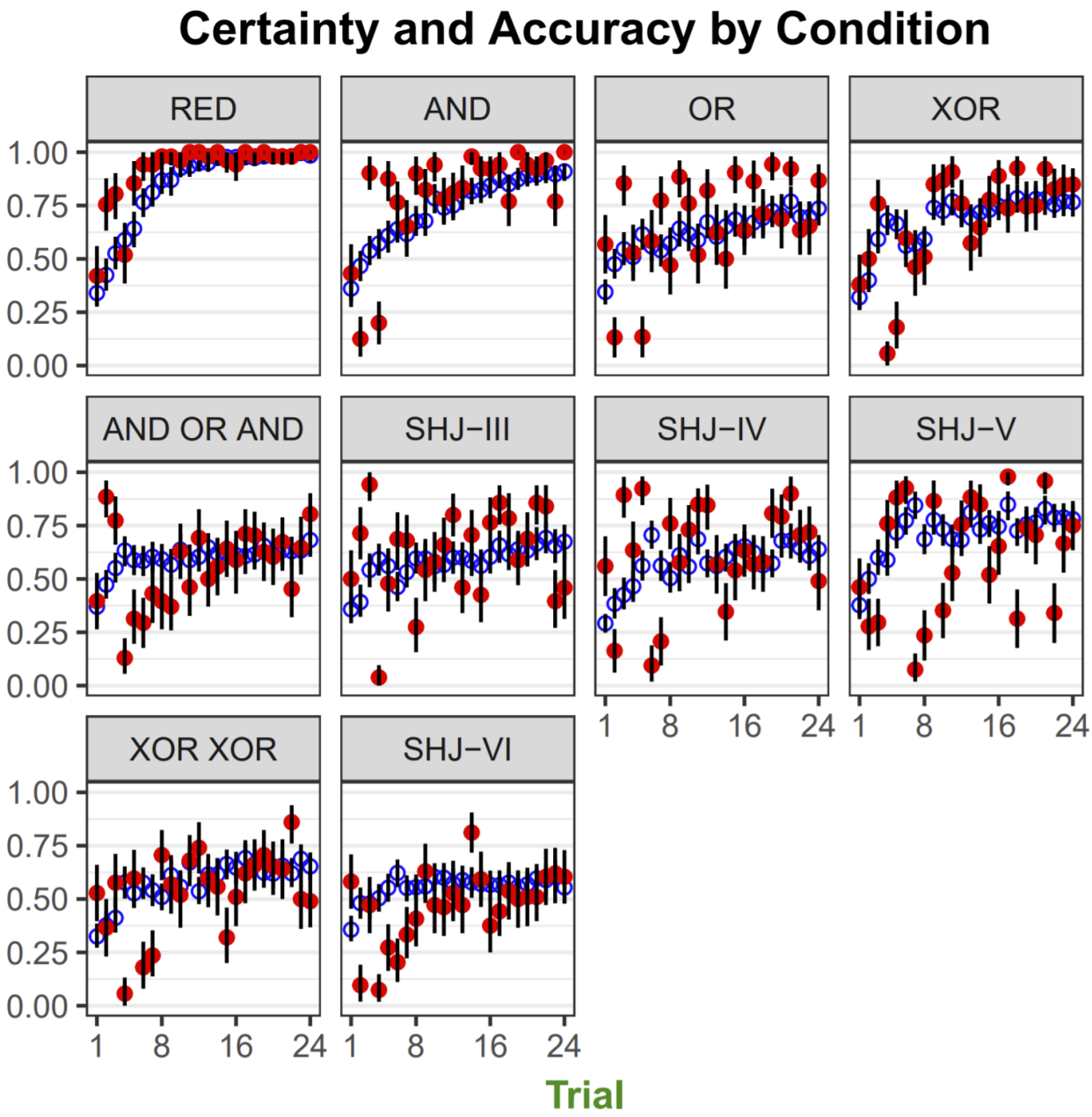

Figure 2-14. Mean certainty (hollow circles) and mean accuracy (filled circles) across concepts for Experiment 4. Chance is 50% across all conditions if guesses are made randomly.

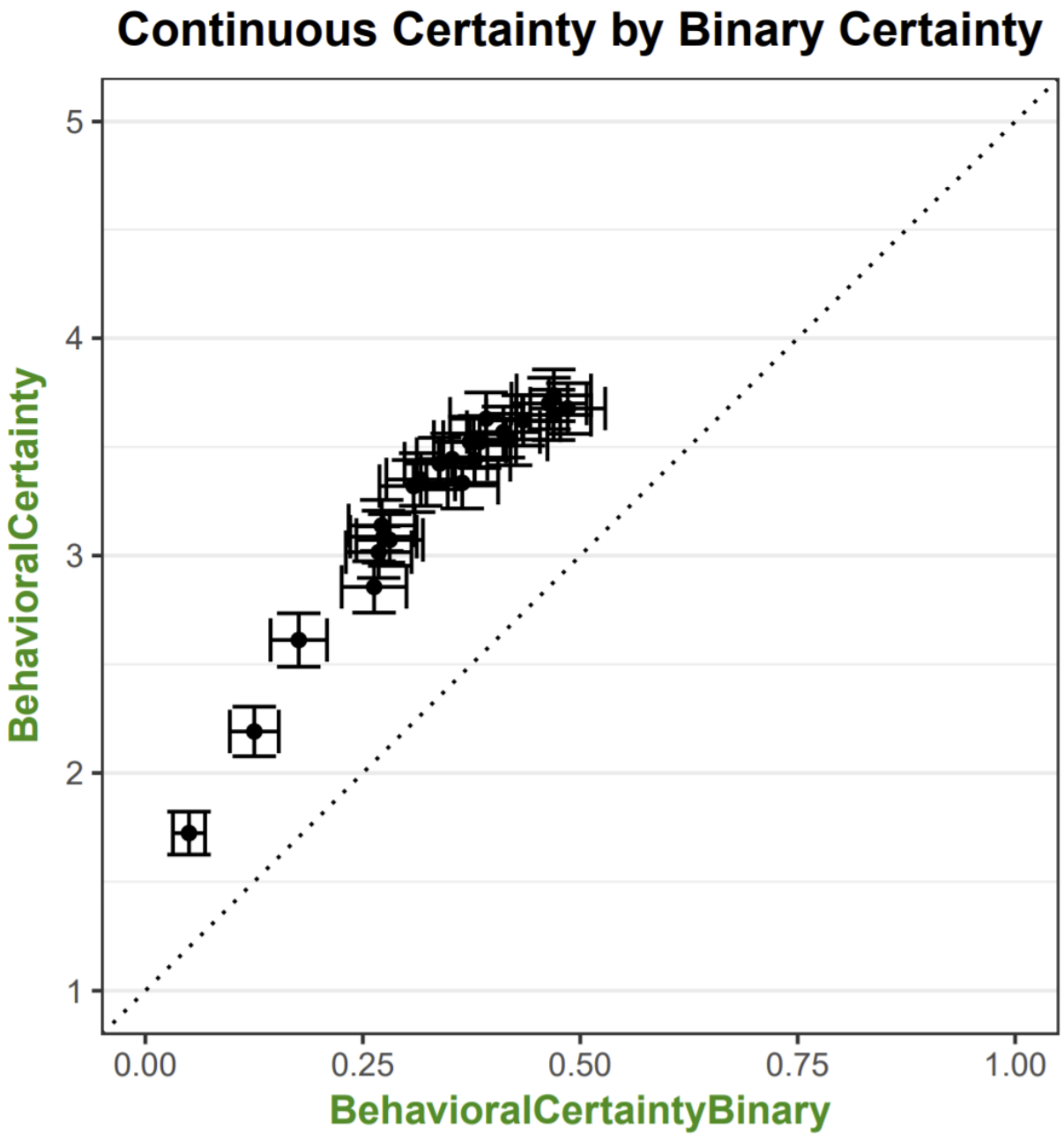

Figure 2-15: Continuous certainty (Experiment 4) and binary certainty (Experiment 1) grouped by trial.

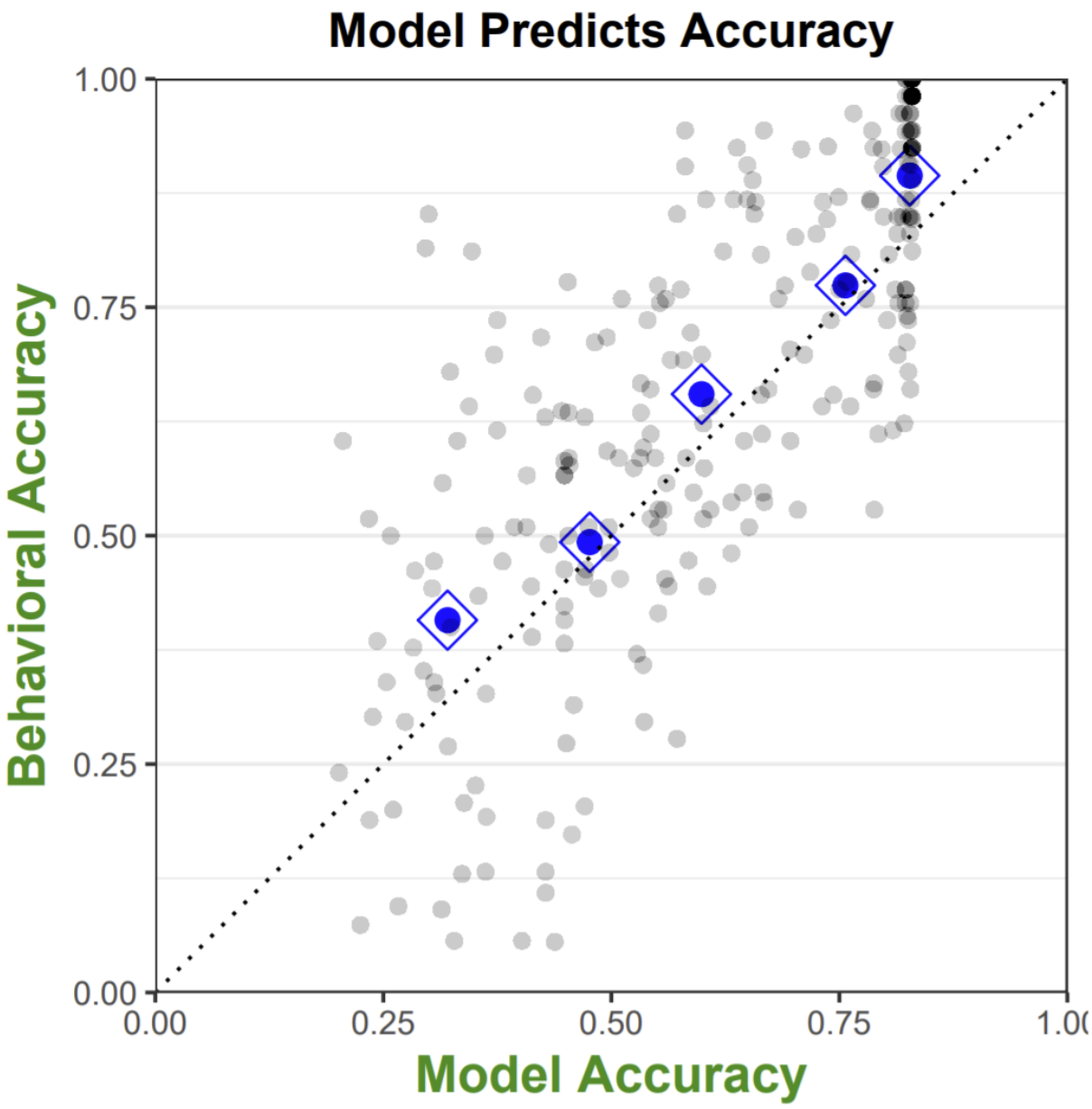

Figure 2-16. Model vs. behavioral accuracy for Experiment 4.

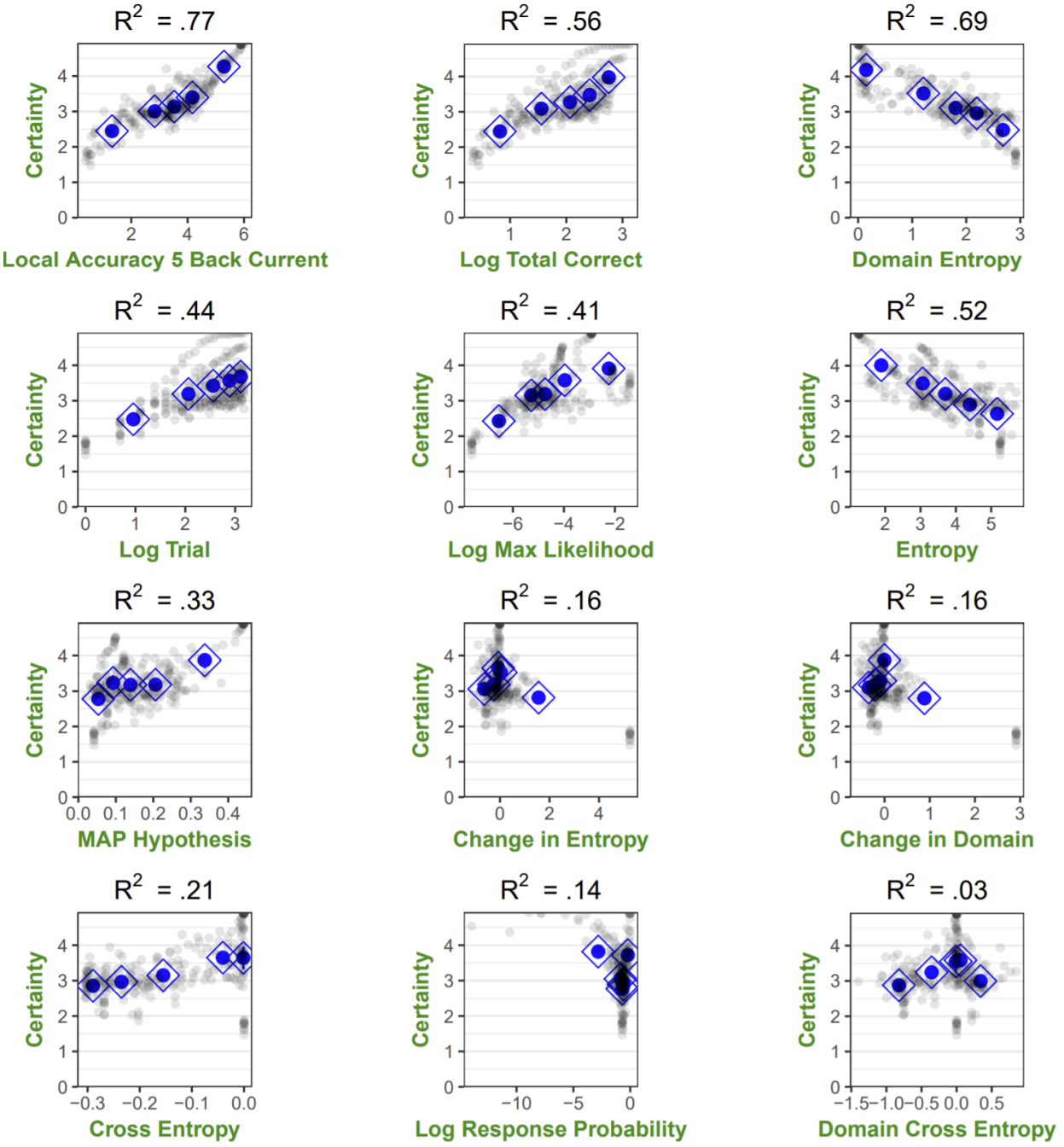

Figure 2-17: Key model fits for Experiment 4.

Chapter III (Latent Diversity in Human Concepts) Appendices

Appendix III-A: Robustness to α

In order to assess the robustness of our general results to the value of α, we also ran our model using alpha values of half (α = .08, reliability = 93%) and double (α = .32, reliability = 80%) the value used in the main results. As Figure 3-8 and Figure 3-9 show, results are not substantially different, providing evidence that our main results are not sensitive to participant reliability. Specifically, our 87% observed reliability yielded 4 - 10 concepts for animals and 7 - 12 for political figures. Increasing reliability to 93% resulted in 5 - 9 for animals and 7 - 11 for politicians. Decreasing reliability to 80% resulted in 3 - 12 for animals and 7 - 12 for politicians.

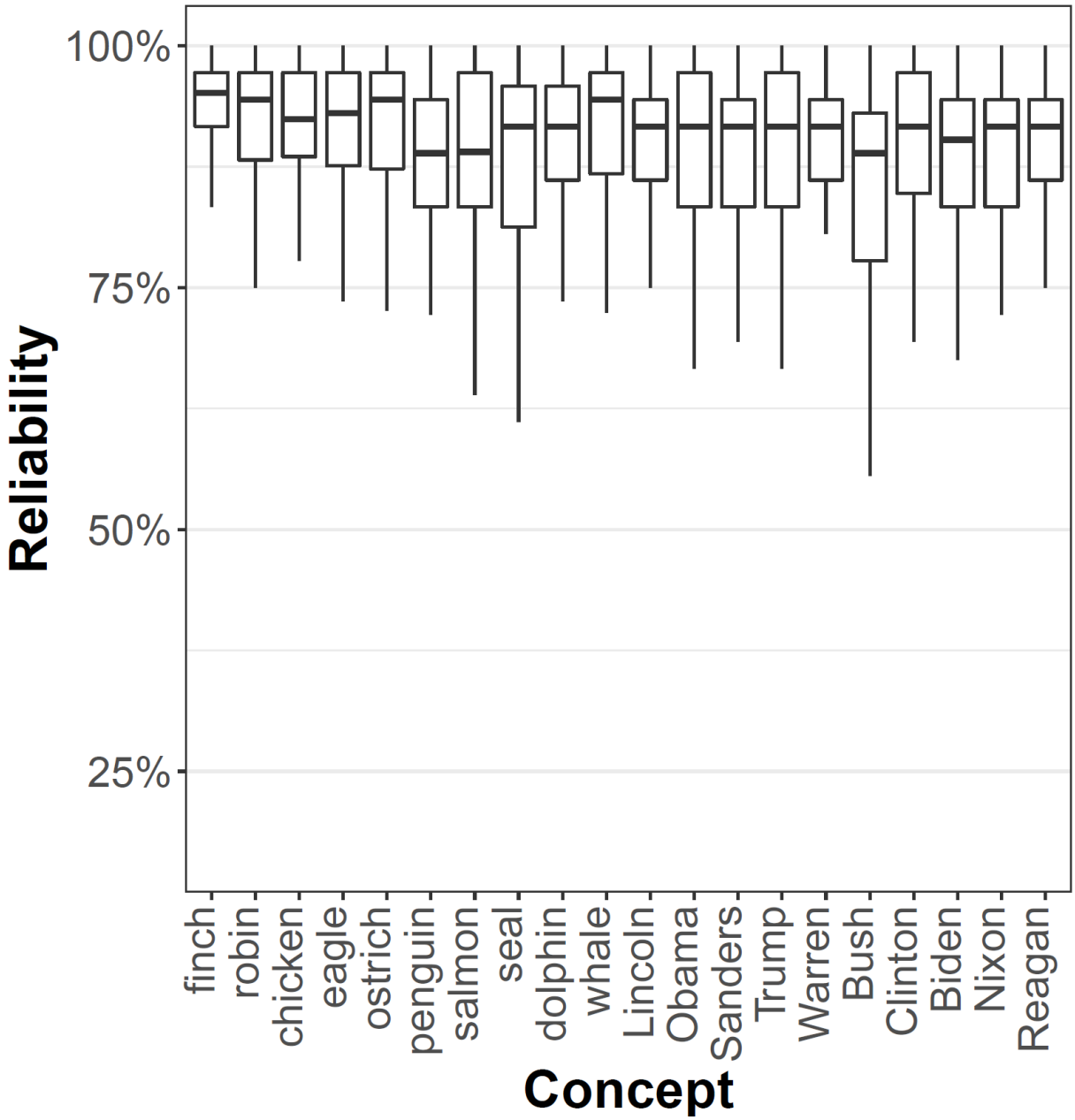

Figure 3-6: Participant reliability scores for each concept in Experiment 1. Boxes show the median 50% reliability quantiles. Median reliabilities for concepts range from 88.8% to 94.4%.

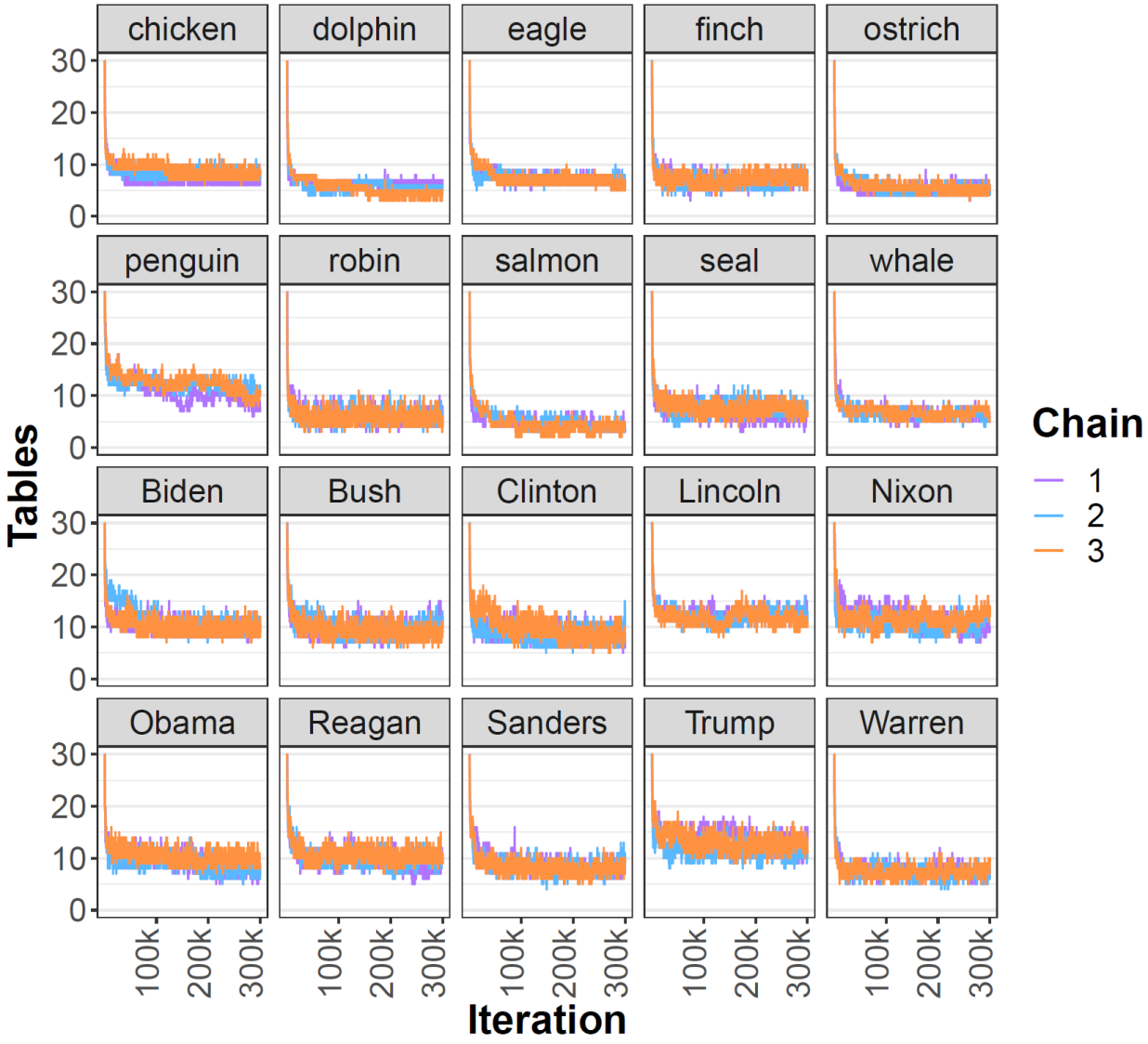

Figure 3-7: Convergence of the clustering model was assessed with multiple runs. Chains converge between 10,000 and 50,000 iterations.

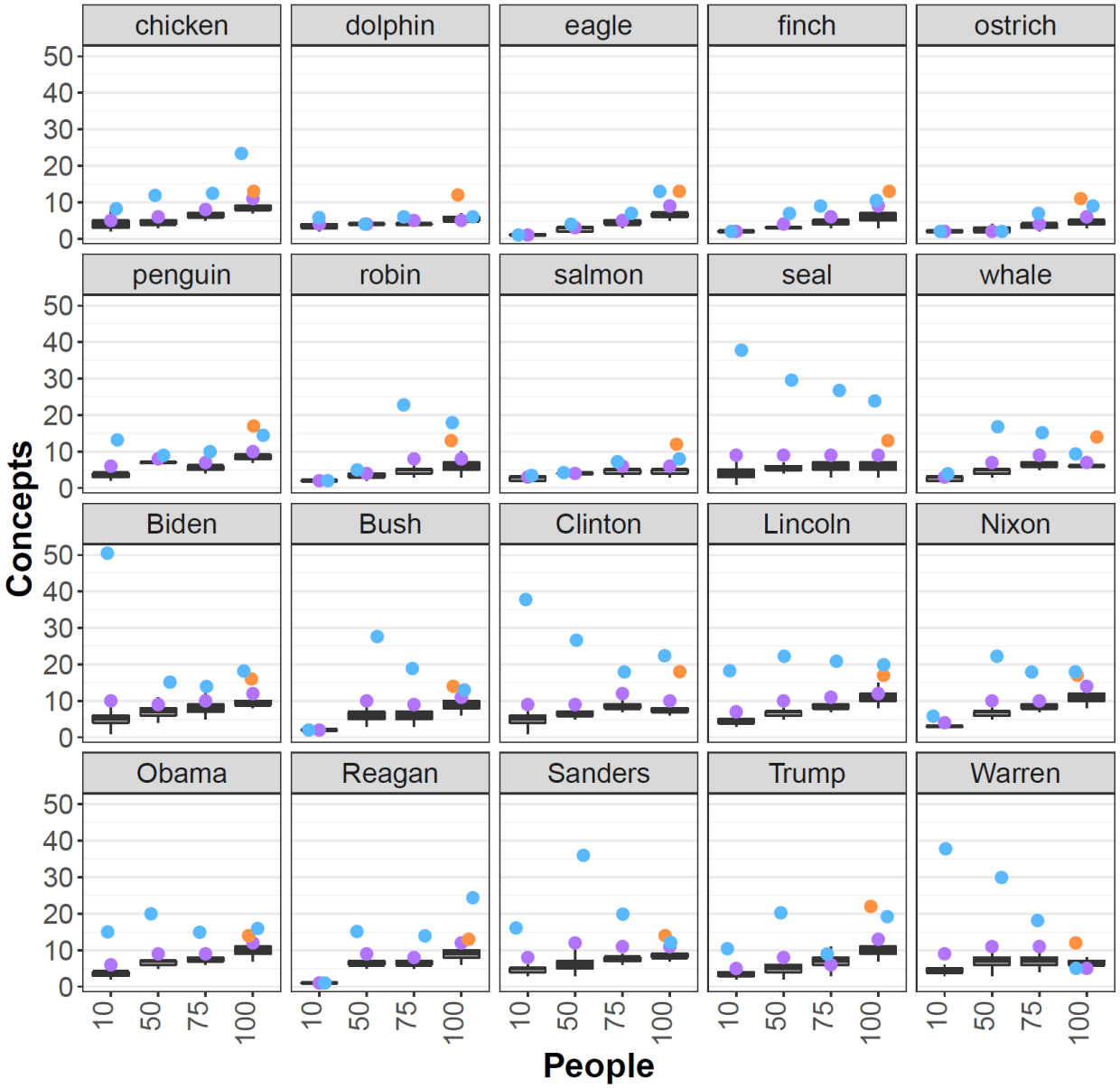

Figure 3-8: Estimated number of concepts (y-axis) depending on the number of people sampled (x-axis) for α = .08, reliability = 93%. Purple data points are the number of clusters for the maximum a posteriori clustering. Orange data points are the number of clusters for the MAP clustering with a uniform prior. Blue data points are a lower bound on the number of concepts estimated by the ecological estimator using the MAP clustering.

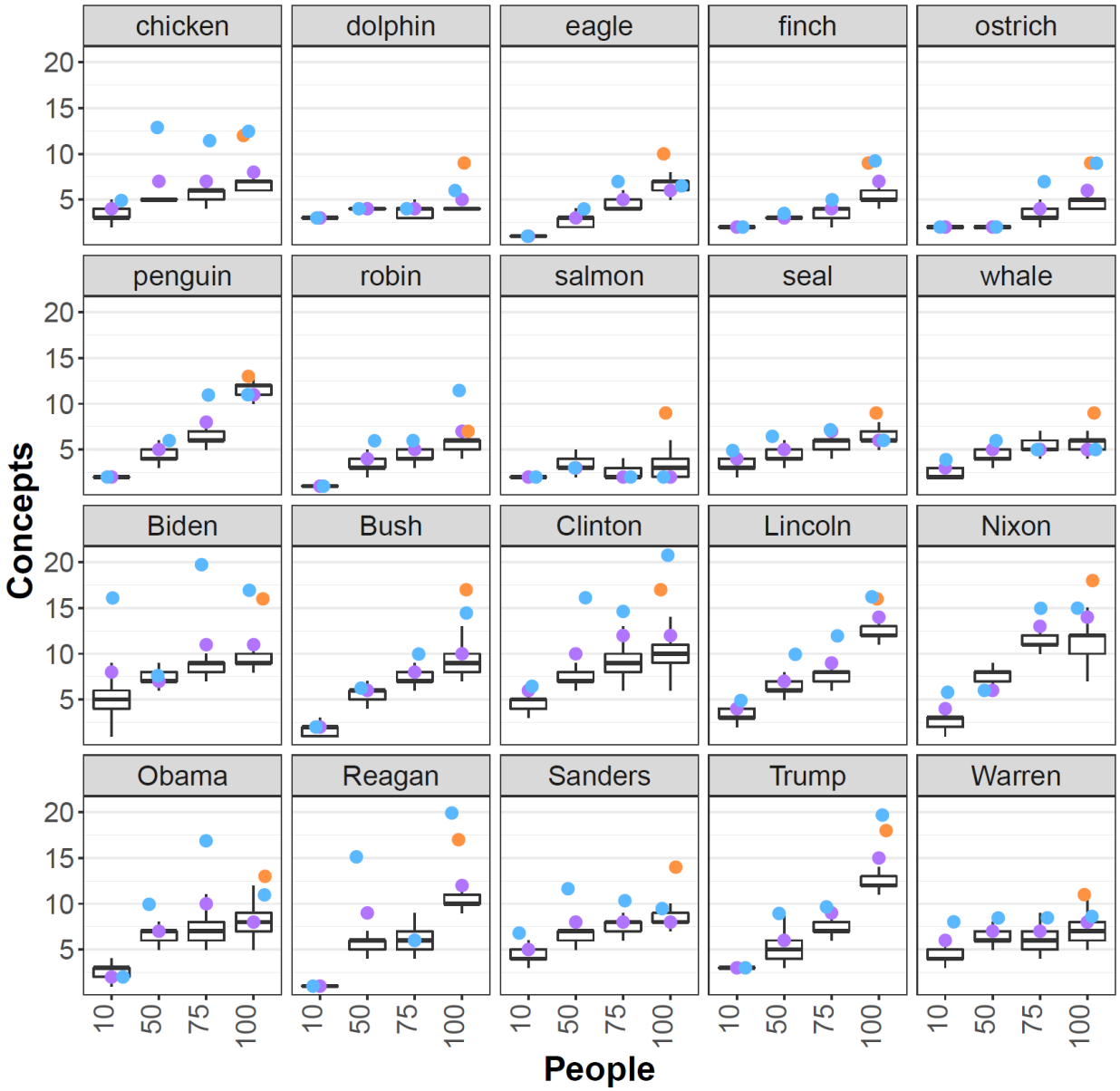

Figure 3-9: Estimated number of concepts (y-axis) depending on the number of people sampled (x-axis) for α = .32, reliability = 80%. Purple data points are the number of clusters for the maximum a posteriori clustering. Orange data points are the number of clusters for the MAP clustering with a uniform prior. Blue data points are a lower bound on the number of concepts estimated by the ecological estimator using the MAP clustering.

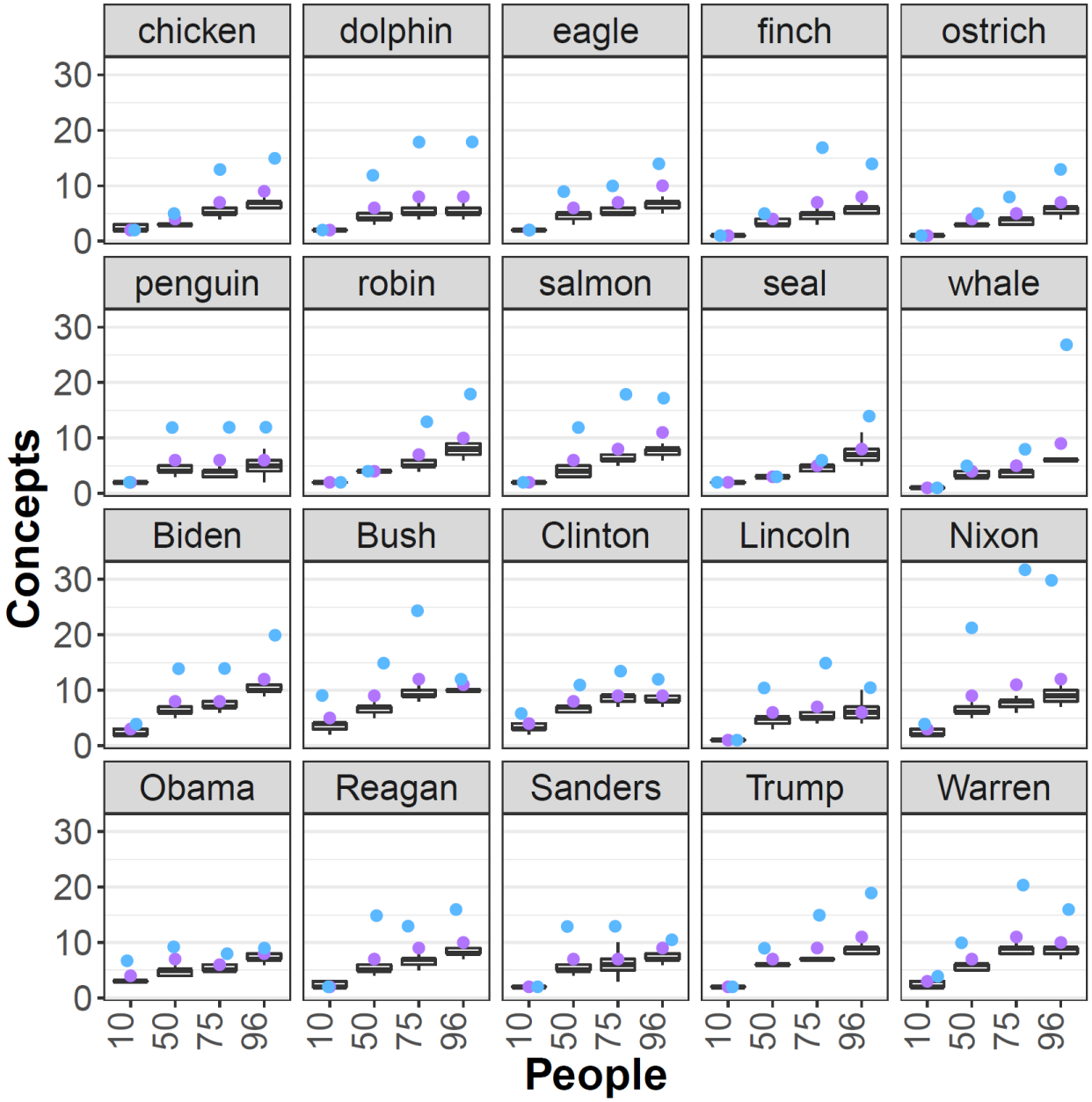

Figure 3-10: Experiment 2 - Estimated number of concepts (y-axis) depending on the number of people sampled (x-axis). Purple data points are the number of clusters for the maximum a posteriori clustering. Blue data points are a lower bound on the number of concepts estimated by the ecological estimator using the MAP clustering.