UNIVERSITY OF CALIFORNIA,
IRVINE

Using Bayesian Cognitive Models in Wisdom of the Crowd Applications

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Irina Danileiko

Dissertation Committee:
Professor Michael D. Lee, Chair
Professor Joachim Vandekerckhove
Professor Mark Steyvers

2018

DEDICATION

To my parents, who encouraged me to pursue math and science.
To Kier, who stuck by and supported me through the program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The research in this document was made possible by all of the guidance and support I received from my advisor, committee members, friends, and colleagues at the University of California, Irvine.

I would especially like to thank my advisor, Michael Lee, for mentoring and supporting me through the graduate program.

This work was supported by NSF Award 1431635.

CURRICULUM VITAE

Irina Danileiko

EDUCATION

Doctor of Philosophy in Psychology **June 2018**
University of California, Irvine *Irvine, CA*

Master of Arts in Psychology **December 2015**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Psychology **June 2013**
University of California, Irvine *Irvine, CA*

Minors: Biomedical Engineering, Linguistics

RESEARCH EXPERIENCE

Graduate Research Assistant **2013–2018**
University of California, Irvine *Irvine, CA*

Visiting Research Assistant **2017**
USC Institute for Creative Technologies *Playa Vista, CA*

Summer School on Computational Modeling of Cognition **2014**
Kapuzinerhof *Laufen, Germany*

Undergraduate Research Assistant **2011–2013**
University of California, Irvine *Irvine, CA*

WORK EXPERIENCE

Graduate Teaching Assistant **2013–2018**
University of California, Irvine *Irvine, CA*

Graduate Student Liaison **2014–2018**
University of California, Irvine *Irvine, CA*

Scientific Consultant for Disabilities Services **2015–2018**
University of California, Irvine *Irvine, CA*

Summer Internship **2017**
USC Institute for Creative Technologies *Playa Vista, CA*

Volunteer Student Teacher **2013**
Third Annual JAGS and WinBUGS Workshop *Amsterdam, Netherlands*

HONORS AND AWARDS

Indiana University, Young Scientist Travel Award	**2016**
A. K. Romney Fellowship in Quantitative Research	**2016**
Society for Mathematical Psychology, Travel Award	**2015**
Women of Mathematical Psychology Travel and Networking Award	**2016**
Society for Mathematical Psychology, Travel Award	**2014**
University of California, Irvine Social Science Merit Fellowship	**2013**
Society for Mathematical Psychology, Travel Award	**2013**
Undergraduate Research Opportunities Program (UROP) grant	**2013**
Summer Undergraduate Research Program (SURP) stipend	**2012**
Department of Cognitive Sciences Honors Program	**2011–2013**

PUBLICATIONS

Danileiko, I., & Lee, M.D. (2018). Using a cognitive model to predict probabilities in sports. In preparation.

Lee, M.D., Vi, J., & **Danileiko, I.** (2018). Testing the ability of the surprisingly popular algorithm to predict the 2017 NBA playoffs. In preparation.

Danileiko, I., & Lee, M.D. (2017). A model-based approach to the wisdom of the crowd in category learning. *Cognitive Science. Accepted 13 Sept. 2017.*

Danileiko, I., & Lee, M.D. (2016). Inferring individual differences between and within exemplar and decision-bound models of categorization. In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Danileiko, I., Lee, M.D., & Kalish, M.L. (2015). A Bayesian latent mixture approach to modeling individual differences in categorization using General Recognition Theory. D.C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Lee, M.D., & **Danileiko, I**. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making, 9(3)*, 259-273.

CONFERENCE TALKS

Annual Meeting of the Cognitive Science Society (2016). Talk title: Inferring individual differences between and within exemplar and decision-bound models of categorization. Philadelphia, PA.

Annual Meeting of the Society for Mathematical Psychology (2016). Talk title: The wisdom of the crowd in category learning. New Brunswick, NJ.

Annual Meeting of the Society for Mathematical Psychology (2015). Talk title: A Bayesian Implementation of General Recognition Theory. Newport Beach, CA.

Applications of Mathematical Psychology in Industry meeting (2015). Talk title: The crowd is wise but biased. Newport Beach, CA.

Annual Meeting of the Society for Mathematical Psychology (2014). Talk title: A wisdom of crowds approach to predicting category structures. Quebec City, Canada.

UCI Undergraduate Research Symposium (2013). Talk title: Calibration of Peoples Estimates of Probabilities in Soccer. Irvine, CA

CONFERENCE POSTERS

Annual Meeting of the Society for Judgment and Decision Making (2017). Poster title: Using the Surprisingly Popular Algorithm to Predict Sports Events. Vancouver, Canada.

Annual Meeting of the Society for Judgment and Decision Making (2016). Poster title: Using a cognitive model to combine probability estimates. Boston, MA.

Annual Meeting of the Cognitive Science Society (2015). Poster title: A Bayesian Latent Mixture Approach to Modeling Individual Differences in Categorization Using General Recognition Theory. Pasadena, CA.

Annual Meeting of the Society for Judgment and Decision Making (2014). Poster title: Aggregating probability judgments using cognitive models. Long Beach, CA.

Summer School on Computational Modeling of Cognition (2014). Poster title: Using the wisdom of crowds to categorize. Laufen, Germany.

Annual Meeting of the Society for Mathematical Psychology (2013). Poster title: Using cognitive models to combine people's estimates of probabilities. Potsdam, Germany.

Annual Meeting of the Society for Mathematical Psychology (2012). Poster title: People's estimation of probabilities in football (soccer) games. Columbus, OH.

ABSTRACT OF THE DISSERTATION

Using Bayesian Cognitive Models in Wisdom of the Crowd Applications

By

Irina Danileiko

Doctor of Philosophy in Psychology

University of California, Irvine, 2018

Professor Michael D. Lee, Chair

The “wisdom of the crowd” phenomenon is when an aggregated group answer to a problem is more accurate than the answer of individuals in the group. Traditionally, aggregation is done with simple statistical methods such as the mean or median of all people’s answers. While these methods can be effective, they don’t allow us to learn about how people make their decisions. Such cognitive processes are not evident from summary statistics. However, we can use cognitive modeling to infer the mechanisms that lead to a person’s decision. In this dissertation, I show how cognitive models can be a powerful tool in accounting for individual differences and biases as well as in helping generate better crowd decisions. I first discuss a hierarchical model for aggregating people’s estimates of probabilities, in an environment for which we know the true answers as well as in a predictive environment for which we don’t know the truth. I demonstrate how using this model, we can identify the experts in the crowd and account for biases in probability perception. I move on to applying the “wisdom of the crowd” idea to the field of category learning. First, I establish that taking the modal response in a categorization task is an effective and accurate crowd measure. I then implement two prominent cognitive models of categorization into a Bayesian framework and apply them to existing category learning data sets. I show how we can learn about individual differences in strategy use and apply these inferences to development of a novel, latent-mixture cognitive model that allows for people to use multiple types of categorization strategies within the same data set. I conclude by discussing the implications of this research for the study of aggregation in the "wisdom of the crowd" effect and in the study of individual differences in decision-making.

Chapter 1

Introduction

The “wisdom of the crowd” is the phenomenon in which an aggregated group answer to a problem is more accurate than the answer of individuals in the group (Surowiecki, 2004). There are at least two ways an aggregate answer can improve upon an individual answer. One way is *signal amplification*, in which combining answers amplifies the common signal and reduces noise. In the standard example of guessing the number of jellybeans in a jar, the idea is that the ground truth of the correct number provides a common signal that some people will reliably detect, while other people might be less consistent in their judgments. The net result is that the group overall will favor an answer near the ground truth, even if some individuals guess numbers very far way. A second way is *jigsaw completion*, in which different individuals solve different parts of the problem. For example, people have varying sets of knowledge and each person might pay attention to different aspects of the jellybean jar. Some people may try to solve the problem by looking at one section of the jar and extrapolating from that while others may try to guess volumes of the jar and of the individual beans. Combining all of the guessers’ sources of knowledge might maximize the accuracy of the group.

Surowiecki (2004) identifies four requirements for a wise crowd. The first is *diversity*: the individuals need to have a range of different opinions and backgrounds. In the jellybean example, there can be individual differences in how each person perceives the jar. In applications of the "wisdom of the crowd" that may involve learning, some people may learn more quickly than others, and some people may achieve eventual levels of categorization accuracy that are higher than other people's. It is also possible that not just the rate and final level of learning will differ, but the nature of the learning itself will differ, with some people learning incrementally and gradually improving their accuracy, and others switching between strategies, leading to sudden changes in accuracy. The second is *decentralization*: the individuals need to draw on different information sources. In the jellybean example, the people might use different estimating strategies. The third is *independence*: the individuals cannot know too much about what others think, so that they provide additional or different information to the group. If the jellybean guessers are asked to generate a guess independently, or are otherwise unaware of the estimates of the other guessers, this requirement will also be met. The fourth requirement is *aggregation*: there must be a method for aggregating individual decisions into a group decision.

Traditionally, simple statistical methods such as the mean or the median has been used to aggregate people's quantitative judgments. For example, for the crowd of people estimating the number of jellybeans in the jar, the group average would be taken as the "wisdom of the crowd" estimate. While these methods can often be accurate, we do not learn about how people make their decisions. One interesting insight would be to identify the experts, or the most accurate, individuals and find out how they generate their decisions. This cognitive process that leads to a person's estimation is not evident simply by looking at the single number they gave as their answer. However, we can infer the mechanisms leading to this decision by using cognitive modeling, which aims to not only describe people's behavior but also predict it. Cognitive modeling is a powerful tool to identify certain biases people may have when it comes to understanding and perceiving decision-making processes (Lee, 2011;

Lee and Wagenmakers, 2013).

The challenge of aggregation has been treated as a cognitive modeling problem in the past (Lee et al., 2011; Merkle and Steyvers, 2011; Turner et al., 2014). The basic data that need to be combined are behavioral observations, generated by cognitive decision-making processes based on people's knowledge. The motivation for a cognitive approach is that it is the knowledge people have, and not their behavioral estimates, that should be combined. This view recognizes that people can be prone to biases and distortions in how they represent and express information. A good cognitive model of their representations and processes can serve to "undo" the distortion, and allow for useful inferences about the knowledge people have. Combining this inferred knowledge can potentially lead to group answers that outperform the statistical combination of the observed behavioral estimates.

This document will explore a few different applications of such cognitive models to environments in which it would aid in a more accurate and informative aggregation process for eliciting a "wisdom of the crowd" effect. In Chapter 2 of this document, I discuss the development of a hierarchical model for aggregating individual behavior and apply it to the problem of combining human estimates of probabilities. This model accounts for biases that people have when estimating probabilities and re-calibrates their estimates. It also allows for individual differences in both this calibration process and in the expertise of individuals. This model is applied to two data sets on existing probabilities of sports and world events. I then discuss an extension to this model for a novel application to a predictive setting in estimating future winning percentages of sports teams. In Chapter 3 of this document, I discuss the application of cognitive modeling to studying the "wisdom of the crowd" in category learning. I show that the "wisdom of the crowd" effect exists in human categorization behavior and move to improve it with cognitive modeling. I review past modeling work in the field with emphasis on two specific, well-known cognitive models of human categorization. I discuss an implementation of one of these models into a Bayesian framework and argue for why such a framework is beneficial. I conclude by demonstrating a novel approach for exploring individual differences in model strategy use and discussing its implications for learning more about human decision-making behavior.

Chapter 2

Aggregating Probability Estimations in Sports Environments

2.1 Introduction

Sports provide an effective real-world statistical environment for studying how people perceive probabilities (Albert et al., 2005; Bar-Hillel et al., 2008). Most people have at the very least basic knowledge of common sports like soccer, baseball, basketball, or American football. Because of this, it is a convenient setting for studying how people make judgments of how often events occur and if they are sensitive to the correct estimations of events that are very rare and conversely very common. This type of quantitative setting also makes certain "wisdom of the crowd" aggregation measures straightforward. For example, when a lot of individuals independently estimate the probability that a specific event occurs in a sports game, we can take the group's average as a crowd measure of that event's probability. This measure can be fairly accurate. However, we do not know anything about how much each individual contributed to that group decision, nor about how each person arrived at their estimate. For this, we need to develop a cognitive model and apply it to real-world statistical events.

In this chapter, I discuss our development of a hierarchical model for the cognitive aggregation of individual behavior to the problem of combining human estimates of probabilities, specifically in sports. I begin by describing its application to existing probability data and then extend it to a predictive setting. Our modeling approach differs in three important ways from previous models that forecast binary events based on combining people's probability estimates (Turner et al., 2014; Ungar et al., 2012). First, our model includes not just calibration processes, but allows for individual differences in both calibration and the expertise of individuals. Secondly, we evaluate the model by collecting data sets in which people are asked to estimate directly the probabilities of events. The detail provided by continuous ground truth probabilities, as opposed to binary outcomes generated from those probabilities, allows for more detailed model evaluation. Thirdly, our modeling approach is completely unsupervised, in the sense that it *never* receives feedback about true probabilities (nor outcomes of probabilistic events generated from those probabilities). We find that our cognitive model for combining people's estimates outperforms simple statistical methods, and that there is interpretable structure about how individuals performed in the inferred individual differences within the model.

2.2 Using Cognitive Modeling to Combine Estimates of Soccer and World Statistics

Turner et al. (2014) use hierarchical Bayesian methods to pursue the problem of using individual judgments to forecast probabilistic events. Their models incorporate a key insight from the existing literature on human estimation of probabilities, which is that people may be miscalibrated in their perception of probabilities (Brenner et al., 1996; Lichtenstein et al., 1982; Yates, 1990). The models developed by Turner et al. (2014) explicitly incorporate calibration processes, and build on the work of Budescu and Johnson (2011) and Merkle (2010) to use hierarchical methods to allow for individual differences in calibration.

A second insight from the existing literature is that there are individual differences in expertise (Weiss and Shanteau, in press). This aspect of individual differences is not included in the models developed by Turner et al. (2014). One way they can be included within a hierarchical modeling approach is developed by Lee et al. (2012), in the context of different but related wisdom of the crowd problem involving ranking data. The basic idea is to assume that people's representations are all centered on a common ground truth, but the expertise of the individual determines how precisely they represent the truth.

2.2.1 Experiments

Survey Setup

We conducted two experiments to collect people's estimates of probabilities in a survey format. The first experiment involved an environment of general knowledge questions, and the second experiment involved an environment of questions relating to soccer games. For both experimental environments, we constructed 40 questions requiring the estimation of a probability or a percentage. For the general knowledge survey, the answers were found from a variety of sources including the 2013 CIA World Factbook, Government websites, the websites of the relevant professional societies, and Wikipedia. For the soccer survey, generating the answers for our questions involved finding and analyzing historical data.

Sports like soccer provide real-world statistical environments that have been widely analyzed (Albert et al., 2005). Because most people have some level of understanding of a popular

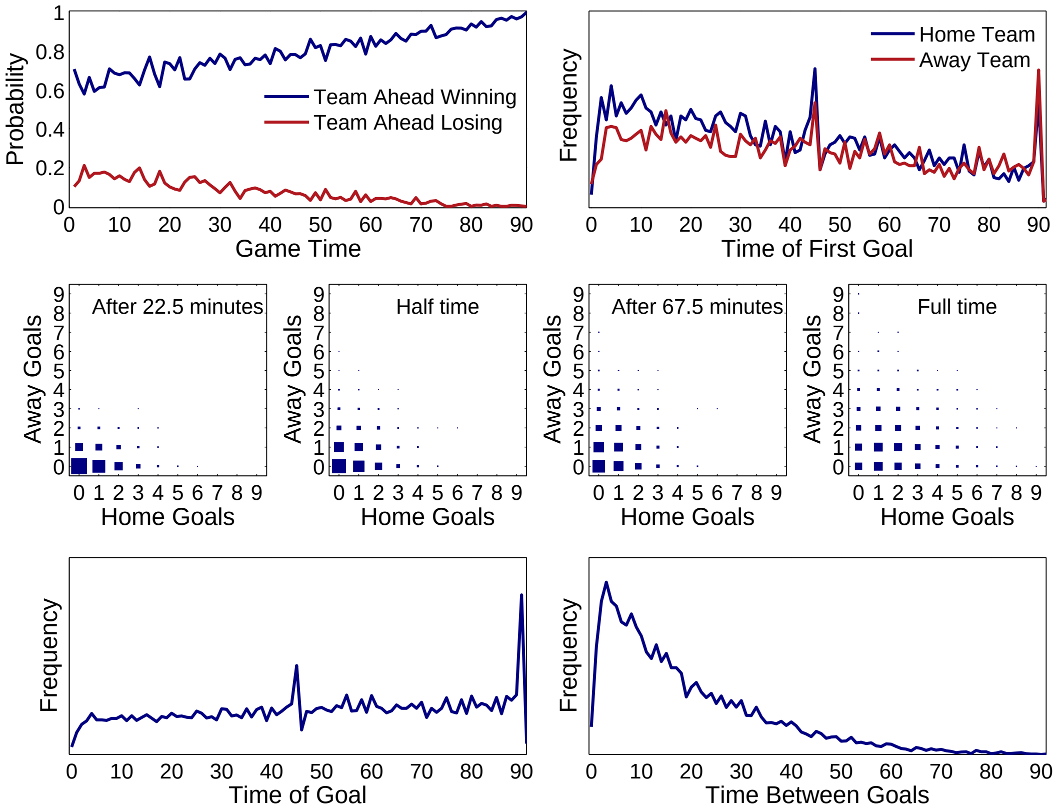
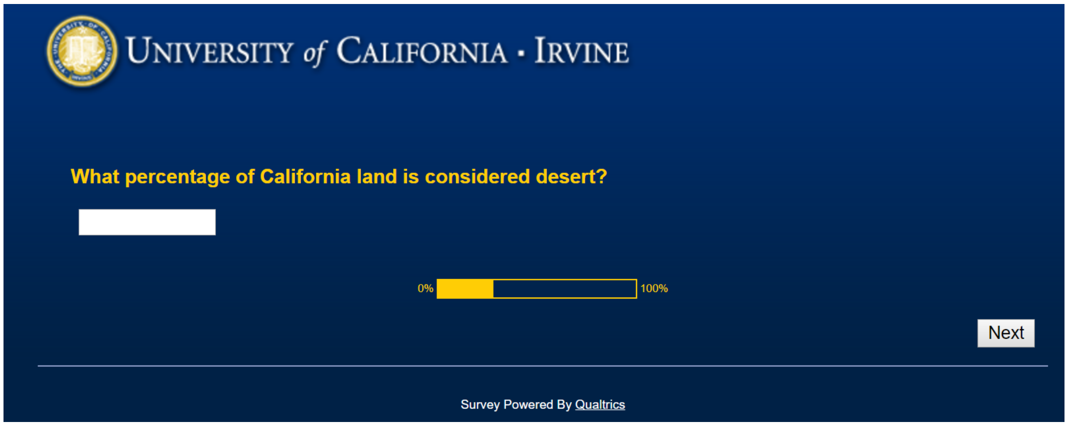


Figure 2.1: Probabilities of events in the soccer game environment, based on 6072 games from first-division games played in domestic leagues between 2001 and 2011.

sport like soccer, and statistics characterizing the outcomes of games are readily available, it is a convenient setting for studying the psychology of probability estimation (Bar-Hillel et al., 2008). To compile the necessary statistical characterization of the soccer environment, the details of 6072 first-division professional games played between 2001 and 2011 in the domestic leagues of a large number of countries were obtained from soccerbot.com. From the information available in these records, we parsed the sequence of goals scored by the home and away team. For example, the information recorded for a US Major League Soccer game between home team Chicago Fire and away team Los Angeles Galaxy was that the home team scored goals in the 1st and 84th minutes, and the away team scored a goal in the 78th minute.

From these goal scoring data, a variety of statistical analyses of the soccer game environment



What percentage of California land is considered desert?

UNIVERSITY of CALIFORNIA · IRVINE
0%
100%
Next
Survey Powered By Qualtrics

Figure 2.2: Screenshot of one sample question from the real-world probabilities survey environment.

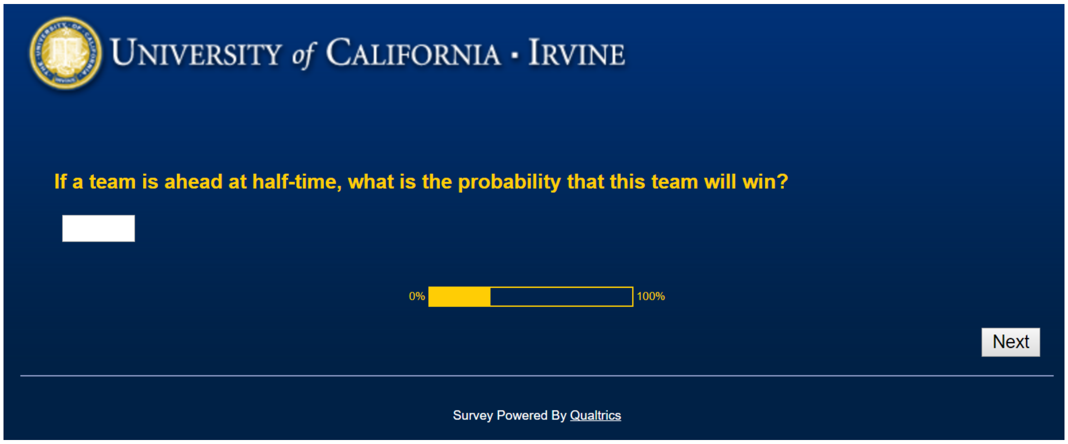
are possible. Figure 2.1 shows a set of analyses on which our probability estimation task questions are based. The panel in the top left shows the (frequentist) probability of the team currently ahead eventually winning or losing a game, as a function of the time at which they are currently ahead. The panel in the top right shows the probability a team will score their first goal at a certain time, for both home and away teams. The sequence of panels in the middle row show the distribution of game scores (i.e., home team goals and away team goals) at four different times during a game. The bottom panels show the distributions of the times goals are scored, and the length of time between goals.

The 40 estimation questions are detailed in Table 2.1, together with the empirical ground truth found by analyses of the game data shown in Figure 2.1. The questions were developed in terms of eight types – such as questions about probabilities of the team ahead winning – with five specific questions for each type. The question types were always completed in the same order listed in Table 2.1, but the order of the specific questions within each type was randomized for each participant. In addition, the questionnaire began by asking participants to self-rate their soccer expertise on a seven-point scale, and answer seven multiple-choice trivia questions involving soccer facts.

We used Qualtrics survey-creation software to generate the surveys. Figures 2.2 and 2.3 show

Question	Answer
What percentage of the world's freshwater is in permanent ice/snow?	69%
What percentage of the United States land is covered by forest?	33%
What percentage of the world's population lives in urban areas?	51%
What percentage of the United States population is between 0 and 64 years of age?	86%
What percentage of the world's population speaks English as a first language?	5%
What percentage of the world's population is between 0 and 24 years of age?	43%
What percentage of the world's electricity does not come from fossil fuels?	33%
What percentage of the world's water is not freshwater?	98%
What percentage of the human body mass is nitrogen?	3%
What percentage of adult human skeleton bones are found in the hands?	26%
What percentage of the United States population has blood type O+?	38%
What percentage of the United States population is of Native American descent?	1%
What percentage of the 2013 congress is women?	19%
What percentage of the United States working population work from home?	9%
What percentage of the United States population between the ages 18 and 44 voted in the 2012 presidential election?	53%
What percentage of the United States population is not foreign-born?	87%
What percentage of the world's population over 65 years of age is women?	56%
What percentage of United States households own a pet?	62%
What percentage of the world's countries are located in North America?	12%
What percentage of coffee beans in the world are produced by Brazil?	30%
Exports make up what percentage of the United States's GDP?	13%
What percentage of London 2012 Olympic medals were won by European countries?	46%
What percentage of the United States population wears glasses (not contacts)?	64%
What percentage of world languages are spoken by more than 100,000 people?	20%
What percentage of FIFA world cups have been won by South American countries?	47%
What percentage of the world population lives on the continent of Asia?	59%
What percentage of NFL teams make it to the playoffs every year?	25%
What percentage of the world's airports are in the United States?	34%
What percentage of the world's landmass is within the United States?	7%
What is America's percentage of world GDP? (2009 - nominal)	25%
What percentage of words in the Oxford English dictionary are verbs?	14%
What percentage of California land is considered desert?	24%
What percentage of American artificial Christmas trees are imported from China?	80%
What percentage of the world's species live in the oceans?	50%
What percentage of the world's protein supply is located in the oceans?	20%
What percentage of the world's energy supply is consumed by Americans?	26%
What percentage of the world's annual petroleum supply is produced by the United States?	6%
What percentage of the United States population lives in counties located on the shoreline?	39%
What percentage of coal consumed in the United States is used to generate electricity?	90%
What percentage of the United States' electricity is generated by wind turbines?	2%

Table 2.1: The 40 general knowledge questions and their answers.



If a team is ahead at half-time, what is the probability that this team will win?

UNIVERSITY of CALIFORNIA • IRVINE
0%
100%
Next
Survey Powered By Qualtrics

Figure 2.3: Screenshot of one sample question from the soccer probabilities survey environment.

screenshots of two sample questions from the two experiments. The 40 total questions were presented in a random order for each participant. After the questions had been answered, each participant was asked a final question "On a scale of 1 (very poor) to 7 (very well), how well do you think you estimated probabilities?"

Participants

For each experiment, a set of 145 participants were recruited using Amazon Mechanical Turk. Participants were paid US$1 for completing the questionnaire within the Qualtrics survey software interface. Completing an experiment took an average of about 20 minutes. Because participants were recruited online, they were not supervised and had access to the internet while completing the estimation task. To address the possibility that participants could search for answers, we vetted questions to insure they could be immediately answered through a simple Google search. This meant that a search using the question text or keywords from the question did not display the answer in the top matches returned by Google visible on the returned page from the search. Of course, participants could have conducted more detailed searches to find the answers, but we could find no evidence for this behavior in the

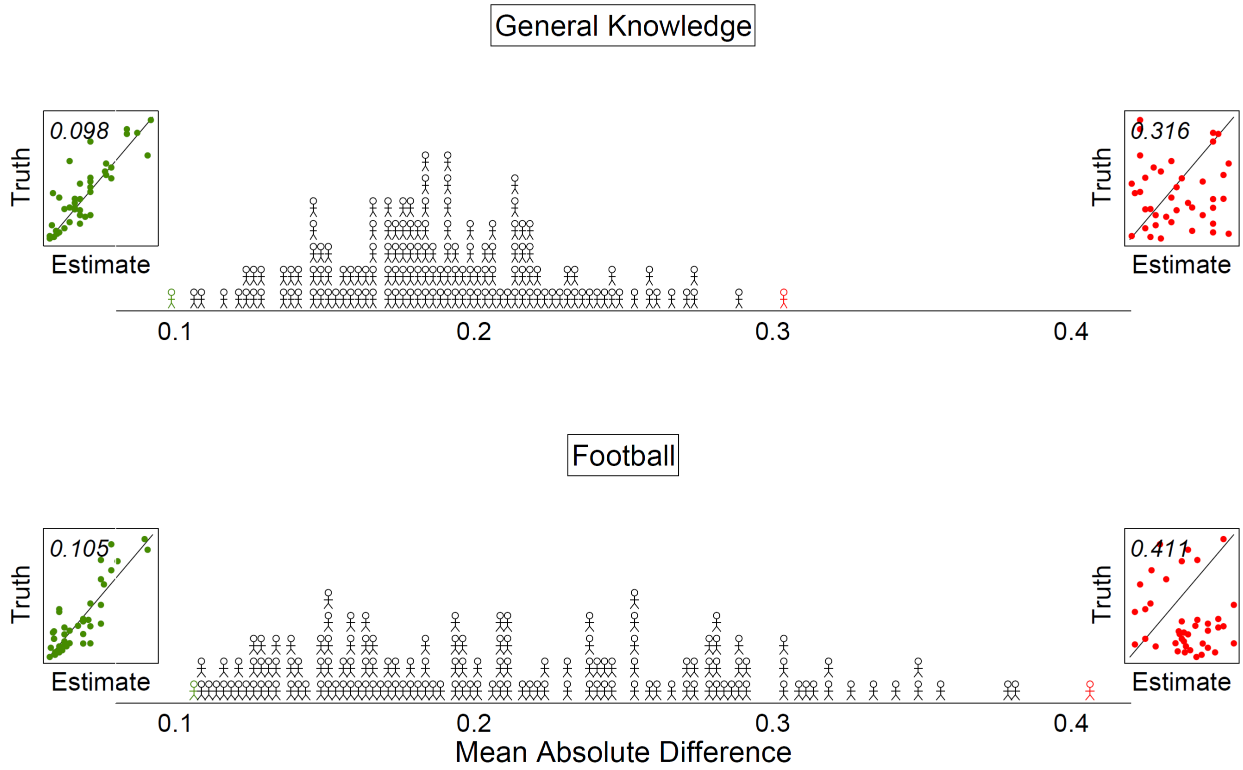


Figure 2.4: Histograms of stick people showing the distribution of performance, measured as the mean absolute difference between estimates and true probabilities, for all participants in both the general knowledge (upper) and soccer (lower) experiments. The inset panels show, for each experiment, the relationship between the estimates and the answers for the best- and worst-performed participants.

accuracies of their answers or the time taken to provide them.

Survey Results

Figure 2.4 summarizes the performance of the individuals in both probability estimation tasks. Performance is measured as the mean absolute difference between a participant's estimates and the answers over all 40 questions. The histograms of stick figures show the distribution of performance for all of the participants in the general knowledge (upper panel) and soccer (lower panel) experiments. It is clear that there is a wide range of performance across people, with the best-performed participants within about 0.1 of the true probabilities on average, and the worst-performed participants 0.3 or 0.4 from the truth on average. Inset with each histogram in Figure 2.4 are two panels showing the performance of the best- and worst-performed participants. These panels show scatter plots of the relationship between the estimates provided by these participants to all of the questions, and the true answers.

2.2.2 Model

Theoretical Assumptions

The primary data collected from each of the experiments consist of probability (percentage) estimates of the 145 participants to all 40 questions. It is straightforward, for each question, to find the mean and median estimate, as standard statistical approaches to combining people's estimates. Developing a cognitive model of the data requires making assumptions about how people represent probabilities, and how they produce estimates. The founding assumption is that the true probability for each question is a latent parameter, represented by π_i for the the ith question. The goal of a cognitive modeling approach is to specify how that knowledge is represented within individuals, and how decision processes act on the knowledge to produce the observed data.

The first psychological assumption involves the miscalibration of probabilities. It has often been found in probability estimation tasks that people systematically over-estimate small probabilities and under-estimate large probabilities (Tversky and Fox, 1995; Gonzalez and Wu, 1999; Zhang and Maloney, 2012). This means that the behavioral estimates generated by people are distorted versions of a person's latent knowledge of the probability. Building a calibration process into a model allows for the distortion to be corrected. The goal is to combine people's latent knowledge, free from the effects of miscalibration. One simple calibration model is shown in the bottom-left panel Figure 2.5 and involves a non-linear function that maps true to perceived probabilities, consistent with the over-estimating small

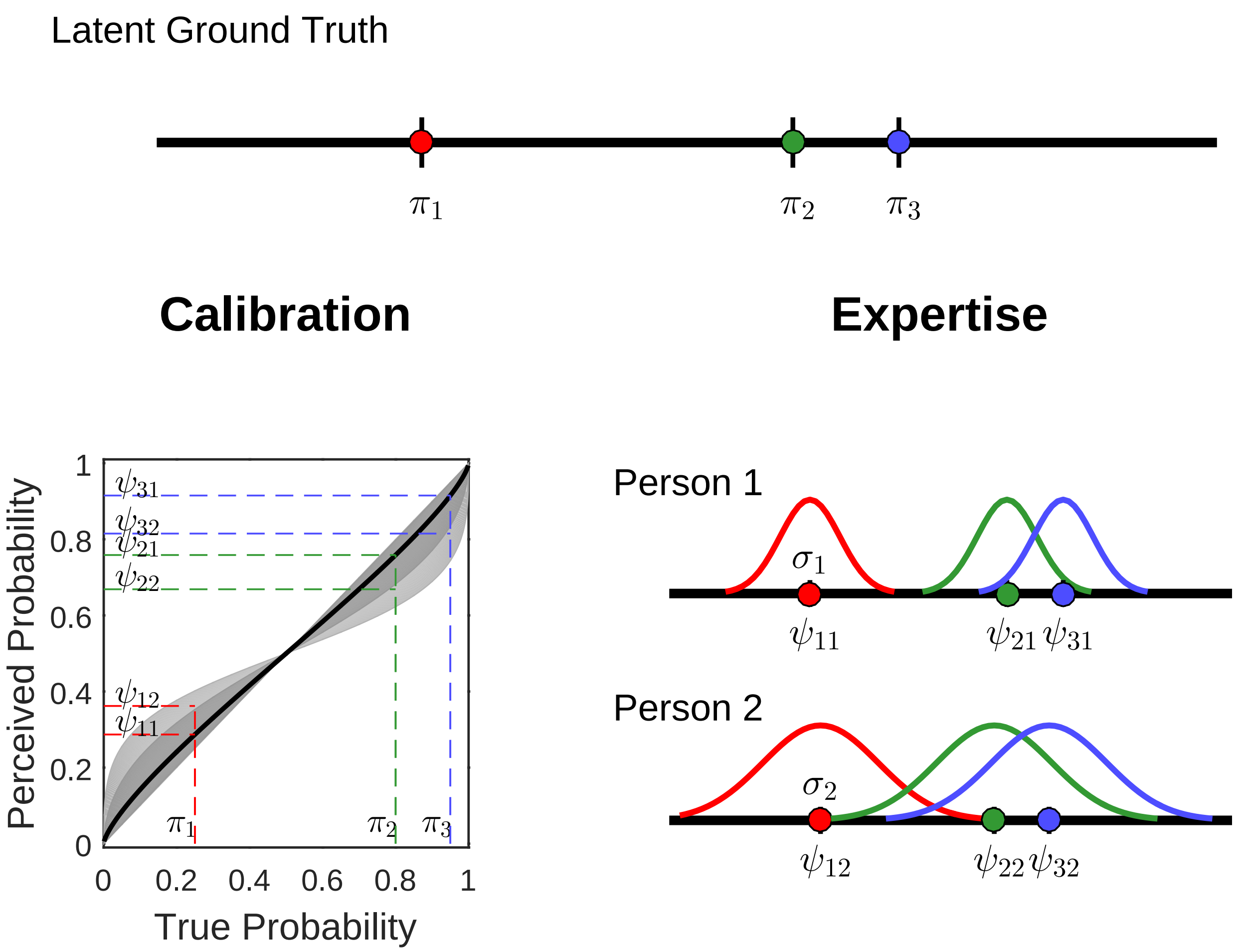


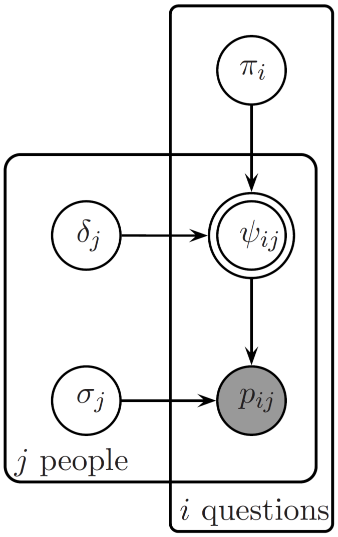
Figure 2.5: The theoretical framework for our cognitive model of probability estimation. The ith probability is assumed to have a latent truth π_i that is subjected to calibration and expertise processes in producing an observed estimate. Calibration operates according to a non-linear function that maps true to perceived probabilities, such that small probabilities are over-estimated and large probabilities are under-estimated. Expertise controls how precisely a perceived probability is reported through the standard deviation of the Gaussian distribution from which the behavioral estimate is sampled. Both the level of calibration and expertise processes are controlled by participant-specific parameters that allow for individual differences.

probabilities and over-estimating large ones. A number of different mathematical forms, motivated in part by different theoretical assumptions, have been proposed for this function, although they share the same basic qualitative properties (Cavagnaro et al., in press; Goldstein and Einhorn, 1987; Gonzalez and Wu, 1999; Prelec, 1998; Turner et al., 2014; Tversky and Kahneman, 1992; Zhang and Maloney, 2012).

We chose to use a linear-in-log-odds functional form with a single parameter capturing the magnitude of over and under-estimation, because it has a natural interpretation that helps in defining the model. The single parameter δ_j for the jth participant scales the log-odds $\log(\pi_i/(1-\pi_i))$ representation of π_i. On the log-odds scale, a probability of $\pi_i = 0.5$ lies at zero and as probabilities move towards zero and one their log-odds representation moves to larger negative and positive numbers, respectively. Thus, scaling a log-odds representation by a factor $0 < \delta_j < 1$ has the effect of "shrinking" a probability towards 0.5. This naturally leads to a transformation that over-estimates small probabilities and under-estimates large probabilities. Thus, the transformed probability on the log-odds scale for the jth participant's perception of the probability for the ith question is given by $\psi_{ij} = \delta_j \log(\pi_i/(1-\pi_i))$.

This calibration function is shown in the bottom-left of Figure 2.5. The expected (mean) prior transformation is shown by the solid line and the 90% and 99% credible regions are shown by successive shading. The transformations of three different true probabilities are shown by three lines, which trace the ith true probability π_i to the perception of that probability by the jth person ψ_{ij}. In the specific examples shown, the first person is well calibrated so ψ_{11}, ψ_{21}, and ψ_{31} are very similar to π_1, π_2, and π_3. The second person, however, is miscalibrated, and so their perceived probability ψ_{12} overestimates the small true probability π_1, while the perceived probabilities ψ_{22} and ψ_{32} underestimate the large true probabilities π_2 and π_3.

The second psychological assumption made by the model involves expertise. Different people seem likely to have different levels of understanding of the true environmental probabilities,



$$\begin{aligned}
\pi_i &\sim \text{Uniform}(0,1)\\
\delta_j &\sim \text{Beta}(5,1)\\
\psi_{ij} &\leftarrow \delta_j \log\left(\tfrac{\pi_i}{1-\pi_i}\right)\\
\sigma_j &\sim \text{Uniform}(0,1)\\
p_{ij} &\sim \text{Gaussian}(\tfrac{\exp(\psi_{ij})}{1+\exp(\psi_{ij})}, \tfrac{1}{\sigma_j^2})
\end{aligned}$$

Figure 2.6: Graphical model for behavioral estimates of probabilities made by a number of participants for a number of questions. The latent true probability π_i for the it question is calibrated according to a parameter δ_j for the jth participant to become the value ψ_{ij}. This calibrated values then produces an observed estimate p_{ij} according to the expertise σ_j of the participant.

and this will affect the precision of their knowledge and the accuracy of their answers. One way to incorporate this assumption, used successfully in a related modeling problem by Lee et al. (2012), is to assume people's estimates are draws from Gaussian distributions that have a level of variability associated with their knowledge. This approach is shown at the bottom right of Figure 2.5. When the jth participant answers the ith question, the assumption is that their estimate comes from a Gaussian distribution that is centered on their perceived probability ψ_{ij}, but has a standard deviation σ_j. The assumption is that σ_j is a property of the participant, and is the same for all of the questions. In this way, the parameter σ_j represents the level of knowledge or expertise of the jth participant, with smaller values corresponding to greater expertise.

Graphical Model

The graphical model in Figure 2.6 formalizes our cognitive model. Graphical models are a standard tool in statistics and machine learning (Jordan, 2004; Koller et al., 2007), and are becoming an increasingly popular approach for implementing and evaluating probabilistic models of cognitive processes (Lee, 2011; Lee and Wagenmakers, 2013; Shiffrin et al., 2008). In graphical models, nodes represent variables and data, and the graph structure is used to indicate dependencies between variables. Continuous variables are represented with circular nodes and discrete variables are represented with square nodes. Observed variables, which are usually data or properties of an experimental design, are shaded and unobserved variables, which are usually model parameters, are not shaded. Plates are square boundaries that enclose subsets of the graph that have independent replications in the model. The attraction of graphical models is that they provide an interpretable and powerful language for expressing probabilistic models of cognitive processes, and can easily be analyzed using modern computational Bayesian methods. In particular, they can be implemented and evaluated in standard software like WinBUGS (Lunn et al., 2000) and JAGS (Plummer, 2003) that automatically approximates the full joint posterior distribution of a model and data.

In Figure 2.6, the underlying latent probability π_i for the ith question is an unobserved and continuous variable, and so is shown as an unshaded circular node. These are the "true" answers to the probability questions that we want to infer. The behavioral data take the form of probability estimates $_{ij}$ given by the jth person for the ith question. These are observed continuous values, and so are shown as shaded circular nodes. The cognitive model describes the process that generates the observed behavior from the assumed latent knowledge. It is important to understand that the model is never provided with the answers to the questions. This means that the latent parameters inferred – the latent ground truths of the questions, and the calibration and expertise parameters of participants – are based solely on using the model to account for the generation of the behavioral data.

The graphical model naturally shows how the two core psychological assumptions convert the latent true probability to the observed behavioral estimate. First, the latent probability π_i is transformed according to the calibration function. Since ψ_{ij} is a function of π_i and δ_j, it is shown as a double-bordered deterministic node. The extent of over- and under-estimation is controlled by the prior distribution of δ_j, which is naturally expressed as a beta distribution. As Figure 2.6 shows, we chose $\delta_j \sim Beta(5, 1)$ because it gives most weight to large values of δ_j that will not transform the latent probabilities drastically, consistent with existing empirical findings and theory. We settled on the exact beta distribution by inspection of the prior distribution for the calibration function it defines, as shown in the bottom-left of Figure 2.5. It is important to note that we defined this prior, as we developed the model, *before* we used the model to analyze data and did not adjust the prior to optimize the results obtained.

The second processing stage produces the estimate p_{ij} as a draw from a Gaussian distribution. The mean is the calibrated probability ψ_{ij} re-expressed on the probability rather than log-odds scale as $exp(\psi_{ij})/(1 + exp(\psi_{ij}))$. The standard deviation of the Gaussian distribution is σ_j for the jth person, and is given a simple weakly informative prior $\sigma_j \sim Uniform(0, 1)$ (Gelman, 2006). The plates in the graphical model in Figure 2.6 replicate over the questions and over the participants. The latent ground truth π_i for each question interacts with the calibration δ_j and expertise σ_j of participants to produce the observed data p_{ij}.

This model is related to the "hierarchical calibrate then average" graphical model presented by Turner et al. (2014), but there are important differences. The Turner et al. (2014) model accounts for the binary outcomes of probabilistic events (e.g., whether a soccer team actually won a game), whereas our model accounts for the underlying probabilities themselves (e.g., the latent probability the team will win the game). Of course, as part of predicting outcomes, the Turner et al. (2014) model determines probabilities that could be assessed against the data from our experiments, and so the difference might be regarded as relatively superficial.

But, it remains the case that these are not the data the model was designed to predict.

More fundamentally, the Turner et al. (2014) model does not incorporate individual differences in the representation of the true probabilities, and uses a different two-parameter form of the linear-in-log-odds calibration function, including an intercept parameter. This difference in modeling assumptions can probably be traced to the thirdand most fundamentaldifference between the two models. The Turner et al. (2014) model *observes* the outcomes of the binary events it is designed to predict, and relies on cross-validation methods for evaluation. Our modeling approach, in contrast, never presents the ground truth probabilities to the model. In machine learning terms, the modeling is fully *unsupervised*, and so models can be directly assessed in terms of their predictions, since there is no possibility of a model being able to over-fit data because of its complexity. This difference requires our model to specify a priori psychologically plausible distributions over models parameters, since they cannot be inferred from data, and so makes the model a more complete attempt to describe the processes involving in people's knowledge of probabilities, their estimation processes, and individual differences in both (Vanpaemel and Lee, 2012).

We also considered reduced versions of our model that included only the calibration or only the expertise assumption. This was done by maintaining either the calibration or the expertise elements of the graphical model in Figure 2.6, but not both, so that only one of the theoretical assumptions was incorporated in the model. Formally, the model that has only calibration used a single σ parameter for all participants, so that $p_{ij} \sim Gaussian(exp(\psi_{ij})/(1 + exp(\psi_{ij})), 1/\sigma^2)$, while the model that uses only individual differences has no calibration function, so that $p_{ij} \sim Gaussian(\pi_i, 1/\sigma_j^2)$. Considering these reduced models allows us to explore whether both calibration and individual differences are useful assumptions, and whether each makes a contribution above and beyond what the other provides.

2.2.3 Modeling Results

We implemented the graphical model in Figure 2.6 using JAGS, and applied it to both the general knowledge and soccer probability estimation data sets. For both analyses we collected eight independent Markov chain Monte Carlo chains, each with 2000 burn-in samples that were discarded and 2000 collected samples. Standard measures of convergence and auto-correlation, including the $\hat{R}$ statistic Gelman (1996), were evaluated to validate the samples as good approximations to the posterior distribution. We implemented and analyzed the reduced models incorporating only calibration or individual differences in exactly the same way.

Model Accuracy

The expectation (mean) of the marginal posterior distribution π_i is a natural measure of a model's inference about the answer to the ith question. These were calculated for the full model, and for the reduced models that included only the calibration or expertise component. In addition, we calculated the mean and the median of the behavioral estimates for each question across all participants as standard statistical wisdom of the crowd estimates.

The performance of each of these five measures – three based on cognitive models, and two on statistical summaries – is shown for the general knowledge experiment in Figure 2.7. The bottom panel shows the distribution of individual participant performance presented in Figure 2.4 and superimposes as vertical lines the performance of the five methods. The best performing method is the cognitive model with calibration and expertise, which produces estimates on average 0.125 different from the true probabilities. The median of participant's answers is 0.127 different on average, followed by the reduced models assuming only calibration or expertise, which are 0.131 different on average. The mean of participant's answers is the worst-performed method, with an average difference of 0.135.

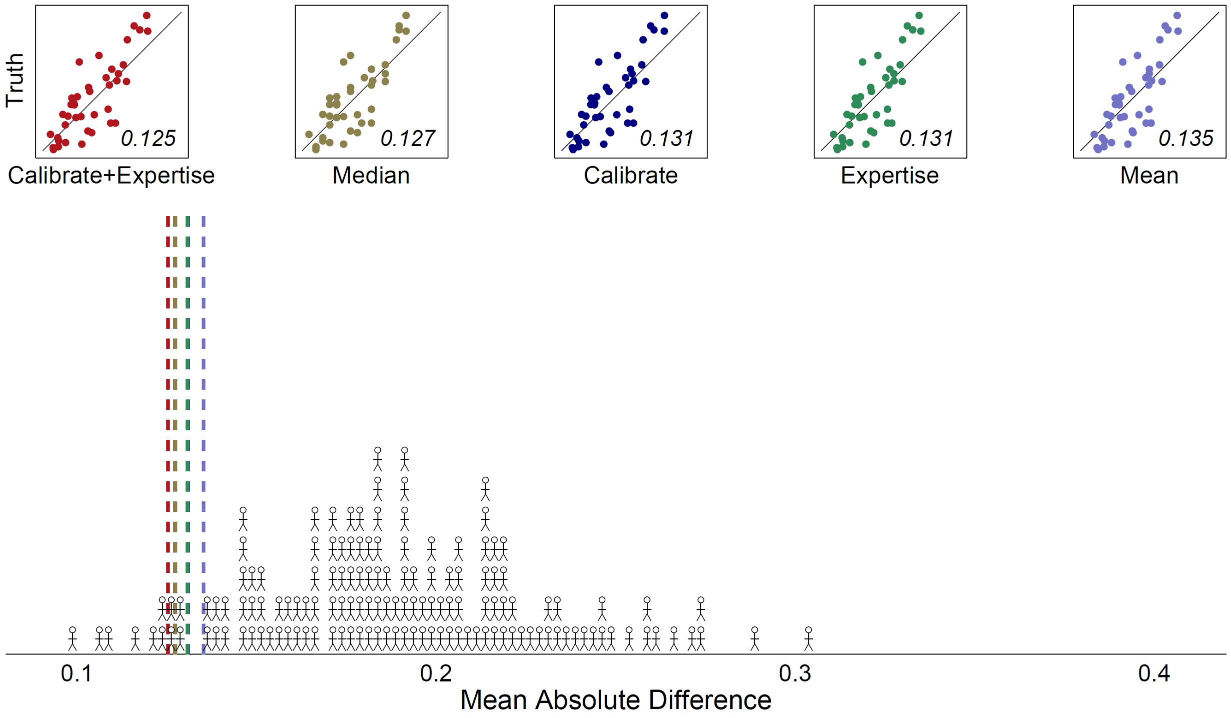


Figure 2.7: The performance of three cognitive models and two statistical methods in estimating probabilities for the general knowledge questions, and the relationship of their levels of performance to individual participants. The cognitive models assume calibration and expertise ("Calibrate + Expertise"), just calibration ("Calibrate") or just expertise ("Expertise"). The statistical methods are the median and the mean of individual responses for each question. The top panels show the relationship between true and estimated answers for all 40 questions for each method. The bottom panel shows the distribution of individual performance as stick figures and the levels of model performance as broken lines. The performance of the models and individuals is measured as mean absolute difference from true answers.

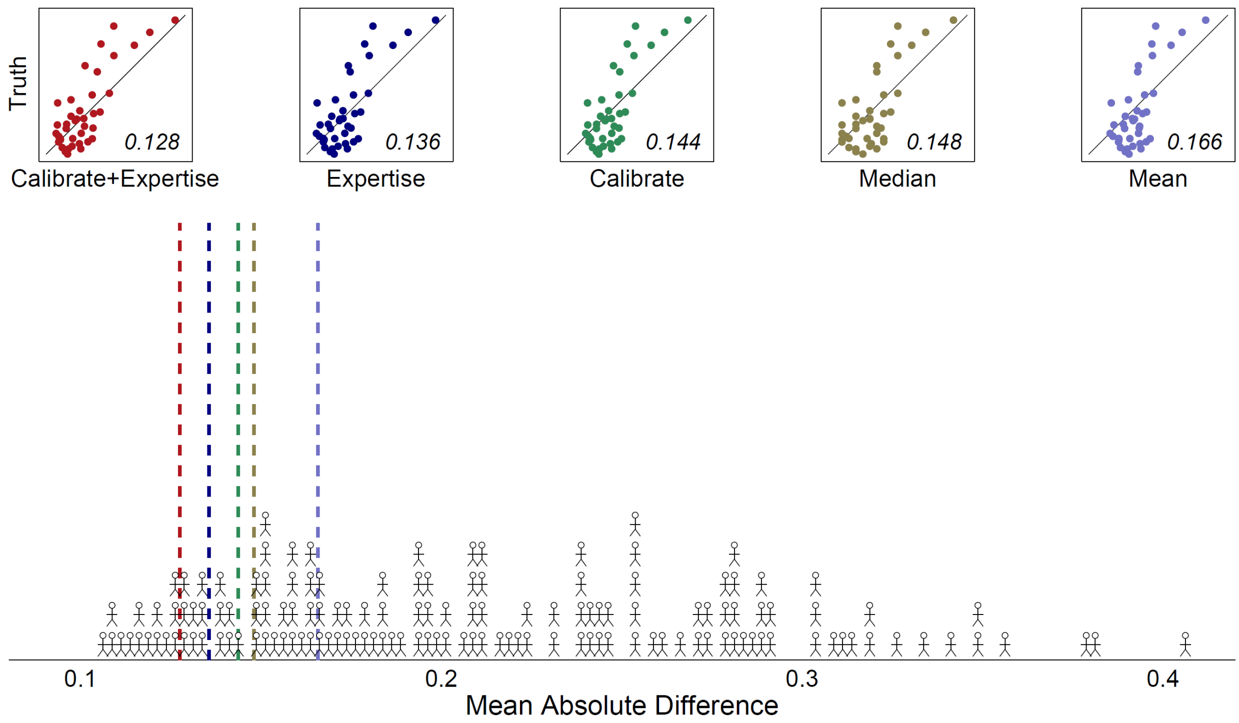


Figure 2.8: The performance of three cognitive models and two statistical methods in estimating probabilities for the soccer questions, and the relationship of their levels of performance to individual participants. The same information is presented in the same format as for the general knowledge questions presented in Figure 2.7.

The inserted panels in Figure 2.7 show as scatter plots the relationship between the true answer and the answer generated by each method. The methods themselves are ordered from left to right from best performing to worst performing, as measured by the average difference between the true answer and the method's answer. It is clear that all of the wisdom of crowds methods perform relatively well, in relation to individual performance, with lower average differences than all but a few individuals.

Figure 2.8 provides the same analysis of estimation accuracy for the soccer questions. The model that includes calibration and expertise performs much better than the other approaches, being an average of 0.128 from the true empirical probabilities, and is again better performed than all but a few individual participants. The two reduced models also outperform both of the statistical approaches. Once again, the best-performed cognitive modeling

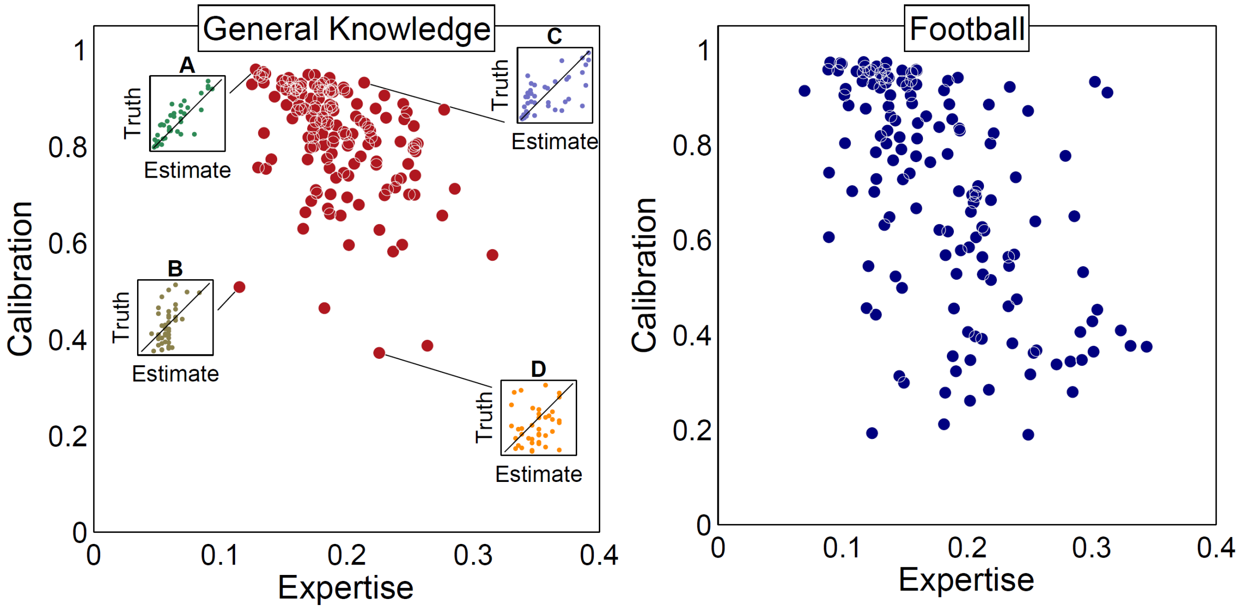


Figure 2.9: The expected posterior expertise σ_j and calibration δ_j parameters for each participant in the general knowledge (left) and soccer (right) experiments. For the general knowledge experiment, four participants are highlighted and the scatter plot of their estimates relative to the answers are shown in inserted panels.

aggregation methods are among the best-performed participants.

We repeated the same analyses using root-mean-squared-error rather than mean absolute deviation as a performance measure for the models and people. All of the important conclusionsthat there are large individual differences in performance, that calibration and expertise is as good as the median and better than all other approaches for the general knowledge data, and better than all approaches for the soccer data, and that the calibration and expertise model performs about as well as the best individuals – all continued to hold.

Individual Differences

The two parameters inferred for each participant are the σ_j measure of expertise and the δ_j measure of calibration. Their expected posterior values for each participant are shown in Figure 2.9 as scatter plots for both the general knowledge (left) and soccer (right) experiments. In both experiments a wide range of values are inferred for both parameters. The expertise parameter – which is a standard deviation for an assumed Gaussian distribution lying on the probability scale from 0 to 1 – ranges from about 0.1 to about 0.3. The calibration parameterwhich is a multiple of log-oddsranges from about 0.95 down to about 0.5 for the general knowledge experiment and lower to 0.3 or 0.4 for the soccer experiment. Thus, it seems clear that both parameters capture variation underlying the probability estimates produced by different participants. There is also no obvious strong correlation between the two parameters for either experiment, suggesting they capture, at least in part, different aspects of the variation in people's performance.

Also shown for the general knowledge experiment in Figure 2.9 are the detailed performance of four individual participants. These participants were selected at the "extremes" of the joint parameter space, to give an indication of the type of estimates inferred to have high and low expertise and strongly or weakly miscalibrated. The participant labeled "A" in the top-left panel is relatively accurate in their estimates and shows no systematic miscalibration, and is inferred to have high expertise and be well calibrated. Participant "B" in the bottom-left panel systematically over-estimates small probabilities but underestimates large ones, and is inferred to be similarly expert but miscalibrated. Participants "C" and "D" show poor performance, and are inferred to be much less expert. Participant "C" does not appear to mis-estimate systematically while the participant "D" does over-estimate small probabilities often. The calibration parameters that the model infers are consistent with this difference.

Figure 2.10 presents an analysis of how the inferred expertise of participants relates to their actual performance on the probability estimation tasks. The top-left panel shows their relationship for the general knowledge questions, while the bottom-left panel shows the relationship for the soccer questions. For both sets of questions there is a strong positive correlation, with people who were inferred to have greater expertise having performed better.

Figure 2.10 also shows how the various self-reported measures of expertise collected in the

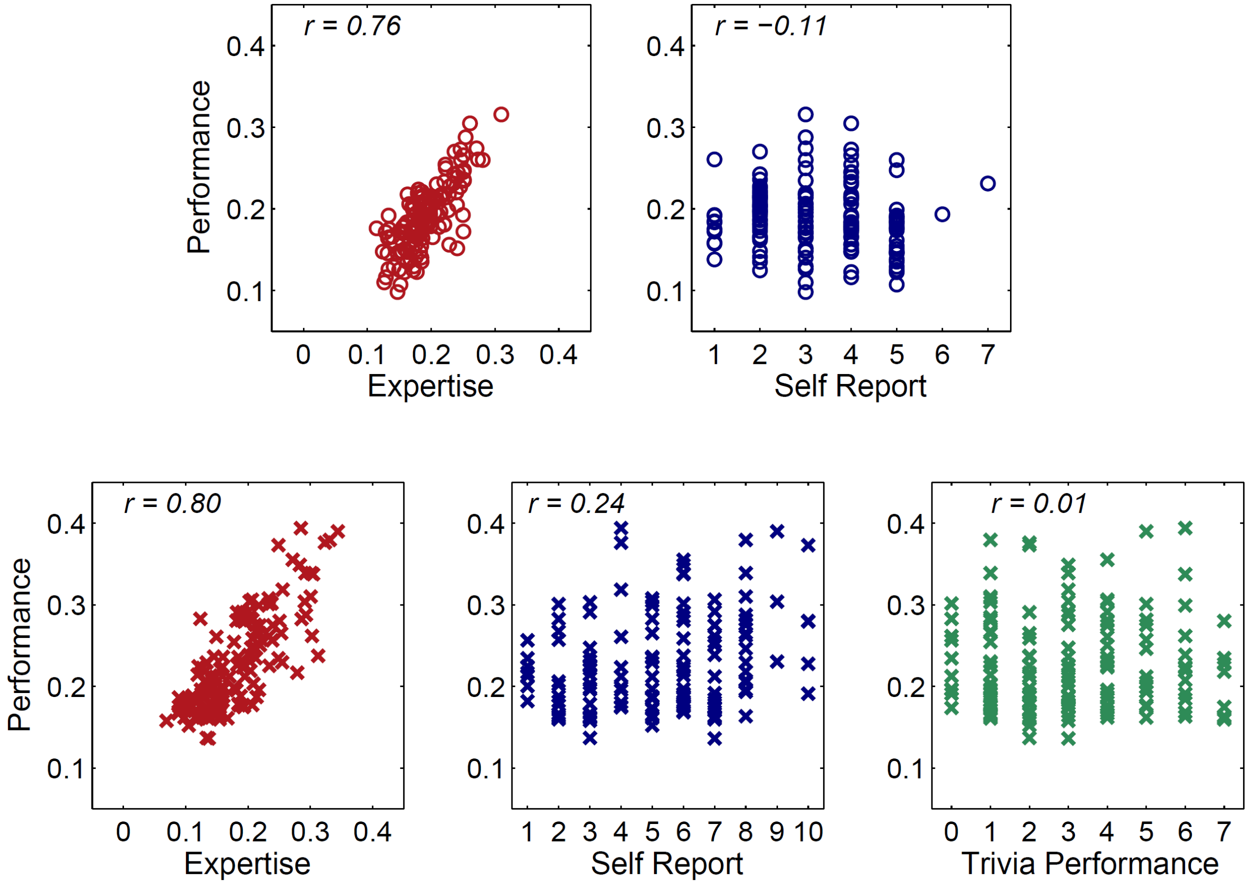


Figure 2.10: The relationship between model-based inferences of individual expertise, self reported measures of expertise, and actual performance in estimating probabilities across individuals. The top two panels relate to the general knowledge questions, and show how the model-based expertise and self reported expertise correlate with performance in estimating probabilities. The bottom three panels relate to the soccer questions, and show how the model-based expertise, self reported expertise, and trivia question performance relate to performance in estimating probabilities. For each scatter plot the Pearson correlation coefficient is also shown.

two experiments relate to performance. The top-right panel shows the relationship between self-reported expertise and performance in the general knowledge experiment. The two panels in the bottom row shows the relationship between self-reported soccer expertise and the number of trivia questions answered correctly in the soccer experiment. None of these self-reported measures are strongly correlated with performance.

2.2.4 Discussion

Our motivating goal was to evaluate a cognitive modeling approach to aggregating probability estimates, and compare its performance to standard statistical methods. Thus, our key result is that a simple cognitive model assuming calibration and individual differences performed as well or better than statistical methods for general knowledge and soccer questions. We built into our model an understanding of the way people often miscalibrate probabilities, as well as an acknowledgment of individual differences in this calibration process, as well as their general expertise. Making these assumptions allowed group answers to be inferred that were as close or closer to the truth than standard wisdom of the crowd methods based on the median and mean. Thus, our results provide some support for the idea that better wisdom of the crowd answers can be found by aggregating over inferred latent knowledge than observed estimates.

As part of formalizing the cognitive process people are assumed to use to make probability estimates, the cognitive model introduces parameters that control individual differences between people. In our model, one process related to calibration and another to expertise, and both involved a single parameter. One perspective on these processes and parameters is that they support the improved inference of the underlying probabilities. Inferring that an individual participant is miscalibrated allows that distortion to be "undone" in the inference of the latent probabilities. Inferring that an individual participant is relatively more expert allows their estimates to be "up-weighted" in inferring the latent probabilities. From this machine learning or statistical perspective, the graphical model in Figure 2.6 can be conceived as a method for the non-linear averaging of behavioral estimates of probabilities that performs well in approximating the underlying true probabilities.

A complementary cognitive science perspective is that the effectiveness of the modeling provides evidence for its core assumptions as important components of human decision-making. From this perspective, the behavioral data collected in the experiments provide further empirical evidence for the systematic miscalibration of probability estimates and for the presence of significant individual differences in expertise. The analyses of the parameters associated with these processes, presented in Figures 2.9 and 2.10, provide insight into basic psychological characteristics of the people providing estimates. The inferred parameter values identify people who are calibrated or miscalibrated and relatively more or less expert.

We emphasize that these inferences are made without knowledge of the answers to the questions. The graphical model in Figure 2.6 does not contain the answers to the questions. This is an especially important point in understanding the contribution the success of our model makes to assessing psychological theory. Leaving the data to be predicted unobserved forces not just the data generating processinvolving calibration and individual differencesto be specified in the model, but also the values of the parameters that control those processes. It is not possible, for example, to infer appropriate values for a calibration function from the performance of the model on previously-made predictions. Instead, the priors for these sorts of parameters must formalize the relevant psychological assumptions, making the model more theoretically complete, and more readily falsifiable (Vanpaemel and Lee, 2012). This means, for example, it is not a trivial result that the model was able to infer the miscalibration of participant "B" in Figure 2.9. It is clear from the scatter plot that this participant over-estimates small probabilities and under-estimates large probabilities systematically, but the model was able to determine this miscalibration without reference to the answers.

Similarly, the ability to infer expertise without reference to the answers means our modeling approach makes predictions about the performance of the individuals. This is because inferred value of the σ_j expertise parameters is available before the answers to the questions are considered. Thus, the strong positive correlation between expertise and accuracy shown in Figure 2.10 has obvious applied possibilities, especially since the basic self-report measures we considered do not correlate performance in the same way. In our experiments, of course, the answers were determined as the questions were generated. But the same modeling approach would apply in genuinely predictive settings for which answers are not yet known. For example, people could be asked to estimate probabilities for soccer games for the upcoming season, so that answers are only available after the season has finished. The model we have developed would immediately make predictions about people's individual expertise and our current results suggest those predictions could be usefully accurate.

It is natural to ask how the modeling approach is able to identify experts and miscalibration without knowing the answers. The easy answer is that it naturally follows from the generative modeling approach we adopted. By specifying a set of reasonable psychological processes as mechanisms for translating latent probabilities to all the observed data, the psychological parameters will be inferred to take the values that make the data likely, and those values should capture psychologically meaningful variation. The beauty of generative statistical modeling is that, having specified how data are produced, the problem of inference is completely and automatically solved by Bayes rule. A more concrete and perhaps more satisfying answer is that the model works by identifying agreement between the participants in the high-dimensional space defined by the 40 questions. Thinking of each person's answer as a point in a 40-dimensional space makes it clear that, if a number of people give similar answers and so correspond to nearby points, it is unlikely to have happened by chance. Instead, these people must be reflecting a common underlying information source. It then follows that a good wisdom of the crowds answer is near this collection of answers, people who are close to the crowd answer are more expert, and people who need systematic distortion of their answers to come close to this point must be miscalibrated. Based on these intuitions, the key requirement for our model to perform well is that a significant number of people give answers that contain the signal of a common information source.

It is also true that the graphical model in Figure 2.6 is not the only possible formalization of the psychological assumptions about calibration and individual differences. Our model assumed that true probabilities were first calibrated, and then subject to individual differences in how accurately they led to behavioral estimates. It would be possible to develop models in which there were first individual differences in the perception of latent true probabilities, and then calibration before estimation. It would also be possible to relax the assumption that each individual has exactly one calibration curve, as parametrized by δ_j, and one level of knowledge, as parametrized by σ_j, that apply to all questions. It would also be possible to introduce different and more general calibration functions, or allows for mixtures of different latent grounds truths, or accommodate a host of more sophisticated psychological assumptions about how people estimate probabilities. All sort of theoretical assumptions are naturally implemented within the graphical modeling framework, using hierarchical extensions and latent mixtures to formalize the processes that generate behavior from knowledge (Lee, 2011).

2.3 Using Cognitive Modeling to Predict Winning Percentages in Sports

The two experiments presented in the above section constitute only a limited test of the model. Each experiment is a significant empirical undertaking – especially the soccer experiment for which answers had to be determined through a time-consuming compilation and analysis of a large dataset – but additional experiments in additional domains should be conducted to test modeling performance further. It would be particularly worthwhile to undertake a genuinely predictive test, asking questions for which answers cannot yet be known.

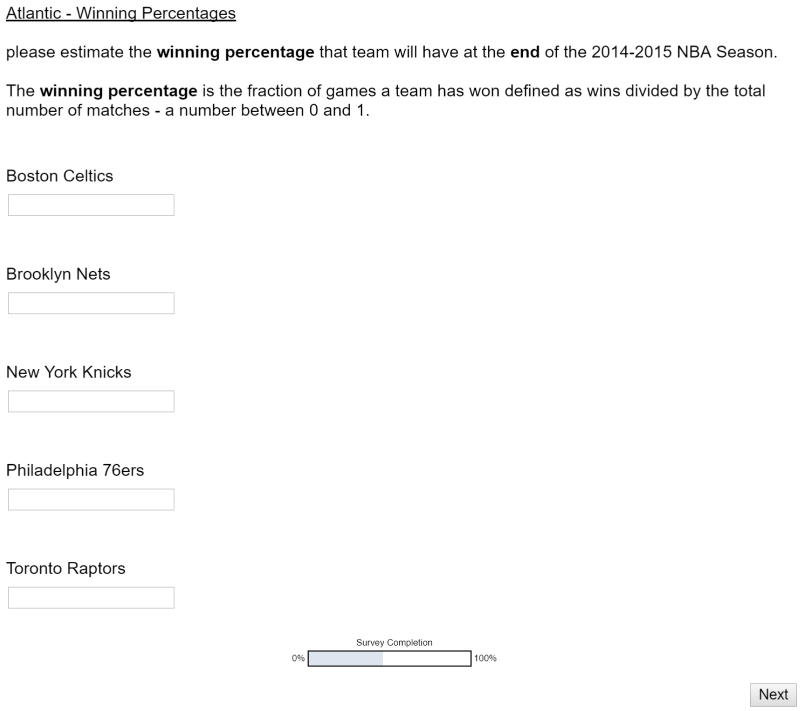
In this section I discuss an extension to our developed hierarchical cognitive model for aggregating people's probability judgments so that it can be applied to the truly predictive setting of estimating winning percentages for teams at the start of sports seasons. To tackle this more challenging problem, we slightly tailor our modeling approach to a predictive setting and evaluate our new model by using it to predict each team's winning percentages in the major US baseball, basketball, and football leagues. We find that our cognitive model performs at least as well as, and some times a little better than, standard statistical aggregation methods. In addition, in contrast with statistical methods, we find our model is able to make accurate and useful inferences about cognitive processes and individual differences.

2.3.1 Experiments

Survey Setup

Our three data sets consider the Major League Baseball (MLB) 2014 season, the National Basketball Association (NBA) 2014-15 season, and the National Football Association (NFL) 2015 season. Final winning percentages for sports teams are reported at the end of the relevant sport season. This percentage is computed by dividing each team's number of wins by the total number of played games in the season. For example, at the end of the 2014 Football season, the Seattle Seahawks had won 12 out of 16 games so their winning percentage was 0.75. For our study, we collected people's estimates of winning percentages in these three sports environments.

The MLB 2014 season, with each team playing 162 games each, started on March 22, 2014



Atlantic - Winning Percentages

please estimate the **winning percentage** that team will have at the **end** of the 2014-2015 NBA Season.

The **winning percentage** is the fraction of games a team has won defined as wins divided by the total number of matches - a number between 0 and 1.

Boston Celtics

Brooklyn Nets

New York Knicks

Philadelphia 76ers

Toronto Raptors

Figure 2.11: Screenshot of a few sample questions from the NBA winning percentages environment.

and data collection happened on May 3, 2014. The NBA 2014-15 season, with each team playing 82 games each, started on October 28, 2014 and data collection happened on October 29, 2014. The NFL 2015 season, with each team playing 16 games each, started on September 10, 2015 and data collection happened on September 9, 2015.

We used Qualtrics survey-creation software to generate the surveys. Figure 2.11 shows a screenshot of a few sample questions from the NBA environment. The MLB and NFL environments looked the same but had questions about teams from those two sports. The questions were presented in a random order for each participant. After the questions had been answered, each participant was asked a final question “On a scale of 1 (very poor) to 7 (very well), how well do you think you estimated the winning percentages?”

Participants

For each of these three sports, a set of 100 participants (99 for the NFL season) were recruited using Amazon Mechanical Turk. For each sport season, participants completed a survey that asked them to predict the final winning percentage of each team at the end of that particular season. For example, a question from the NFL 2015 survey would be “Please estimate the winning percentage that the Denver Broncos will have at the end of the 2015 NFL Season.” This meant 30 questions, or number of teams, in the MLB and NBA seasons and 32 questions in the NFL season. The participants were paid US$1 for their completion of the survey.

2.3.2 Model Extension

Our original model made simple assumptions about how people miscalibrate probabilities and differ in expertise. Miscalibration was modeled using a standard linear-in-log-odds function with a single parameter for each individual, corresponding to the severity of their miscalibration. Lack of expertise was modeled as the addition of Gaussian noise to miscalibrated probability estimates, with another single parameter for each individual, corresponding to the level of the noise. The current model extends on the original model in a number of ways, which I describe conceptually in the remainder of this section, before presenting the model formally in the next section.

The model developed here involves two adjustments to the original model to make it more suitable for the predictive task in these sports domains. The first adjustment is to the prior knowledge of how mis-calibrated people are in their percentage judgment. The second adjustment is to the prior knowledge of people’s expertise in the different domains. In the next two sections, I describe both of these key psychological components. The third section provides formal details of their computational implementation in our cognitive model.

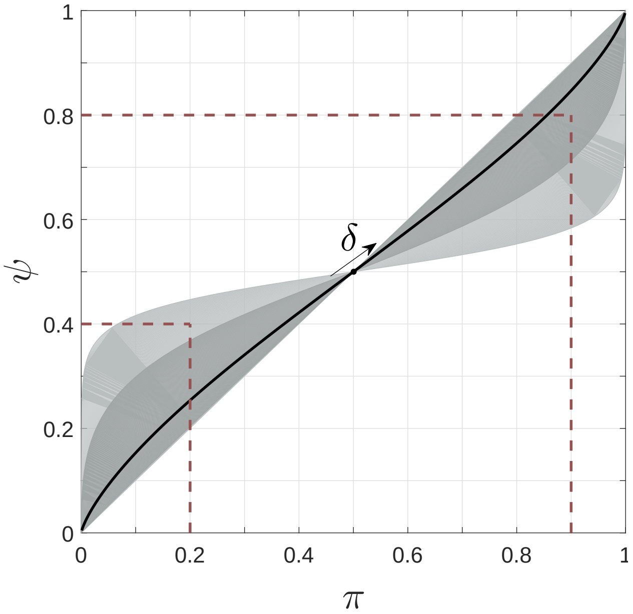


Figure 2.12: Theoretical framework of how probability calibration is represented in our model. The red dashed lines show how a person's representation ψ might be different from the latent truth π.

Theoretical Assumptions

There is a body of literature that has found that people have a tendency to systematically under-estimate large probabilities and over-estimate small probabilities (Tversky and Fox, 1995; Gonzalez and Wu, 1999; Zhang and Maloney, 2012). Additionally, people violate rules of probability when judging events, which results in super- and sub-additivity. In order to get a more accurate representation of a probability estimate, we implement a calibration process into the model that corrects these probability distortions. Figure 2.12 shows a calibration framework that uses a non-linear function to map true probabilities to perceived ones. The function we chose is a linear-in-log-odds function that has two parameters. The parameter δ controls this degree of under- and over-estimation.

We assume that there can be individual differences in the shape of the calibration curve so each person j has a δ_j parameter that govern how they represent probabilities. Each person has a probability estimation that is then scaled according to the calibration function

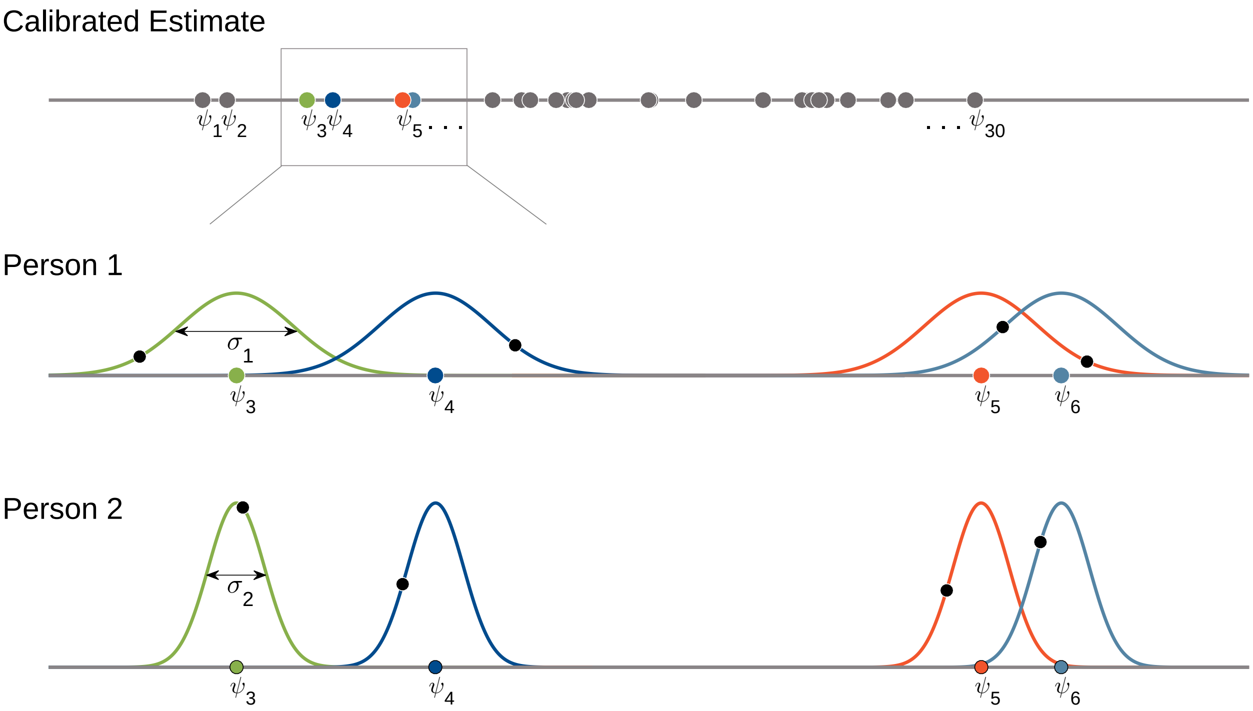


Figure 2.13: Theoretical framework of how expertise is represented in our model as standard deviation of the probability estimation distribution, showing how person 2 would be modeled as more of an expert than person 1.

seen in Figure 2.12. The entire range of probabilities is affected by moving estimates away from the extremes, which results in an under-estimation of large probabilities and an over-estimation of small probabilities. An example of how this may affect a person's estimate is shown by the red dashed lines in Figure 2.12, with over-estimation of $\pi = 0.2$ as $\psi = 0.4$ and under-estimation of $\pi = 0.9$ as $\psi = 0.8$.

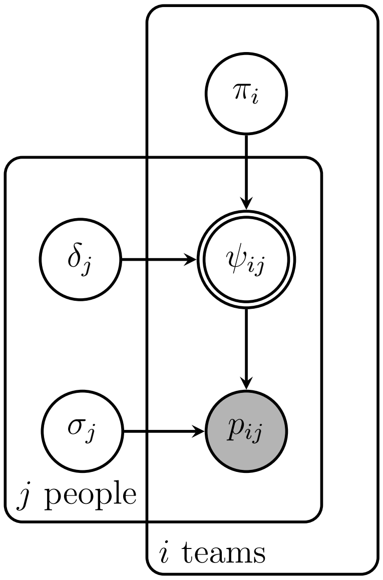
Different people have different levels of knowledge of the chosen sports environments. One person may have been following the sport for a long time and know the distributions of winning percentages while another may be a new viewer and not know how teams perform. In our wisdom of the crowd approach, we want our model to identify and up-weigh the experts.

Figure 2.13 shows one way to implement this idea, previously used in similar modeling situations (Lee et al., 2012; Lee and Danileiko, 2014). Each of the 30 items in the questionnaire has a latent ground truth π_i, that has been adjusted according to our calibration function. We assume each person j draws this re-calibrated probability judgment ψ_{ij}, shown in the bottom two panels of the figure with black markers, from a Gaussian distribution that is centered on ψ_{ij} and with some standard deviation σ_j. This standard deviation corresponds to the expertise level of that person. The smaller σ_j is, the higher that person's expertise because they are drawing estimations closer to the latent truth. We assume that this level stays constant for all questions.

In this case, person 2 would be more of an expert than person 1. As a result of a wide distribution, when attempting to represent a probability, person 1 is more likely to draw their estimate further away from the estimation ψ_{ij}. This may lead to errors in judgment. For example, person 1's distribution is so wide that when estimating the probability for the fifth question, they draw an estimate that is instead closer to the latent truth for the sixth question. Person 2, however, has a more narrow distribution so draws an estimate for the fifth question that is closer to the latent truth.

Graphical Model

We implemented our cognitive model as a graphical model, seen in Figure 2.14. Graphical models have been used as statistical and machine learning tools (Koller et al., 2007). More recently they have also provided a useful and straightforward way to implement probabilistic models of cognitive processes (Lee and Wagenmakers, 2013; Lee, 2011, in press). Parameters and data are represented by nodes and the cognitive processes that generate data are formalized by the structure of the graphical model. Latent parameters are represented as unshaded nodes, while observed data or parameter values are represented as shaded nodes. Continuous parameter or data values are represented as circular nodes, while discrete values are represented by square nodes. Deterministic values are represented by double-bordered nodes. Rectangular plates surrounding parts of the graphical model indicate replications of



$$\begin{aligned}
\pi_i &\sim \text{Uniform}\big(0,1\big)\\
\delta_j &\sim \text{Gaussian}\big(1,\tfrac{1}{(0.5)^2}\big)T\big(0,\big)\\
\psi_{ij} &= \delta_j \log\big(\tfrac{\pi_i}{1-\pi_i}\big)\\
\sigma_j &\sim \text{Exponential}\big(1\big)T\big(0.01,\big)\\
p_{ij} &\sim \text{Gaussian}\big(\tfrac{\exp(\psi_{ij})}{1+\exp(\psi_{ij})},\tfrac{1}{\sigma_j^2}\big)
\end{aligned}$$

Figure 2.14: Graphical model representation as applied to the individual probability judgment behavior.

that structure within the model. This type of model formalism allows us to do fully Bayesian inference using Gibbs sampling programs like JAGS (Plummer, 2003).

In Figure 2.14, the latent ground truth probability for each question i is represented by the node π_i. This is what the model infers as the predictive winning percentage for each team in the sport environment. It is important to note that since we ran the questionnaires and model inferences before the start of each respective sport season, the model is never provided with the final winning percentages, or the question answers. All of the latent parameters of the model, such as the individual-specific calibration curves and expertise values, are inferred by the model's account of how the behavioral data was generated.

The probability π_i gets transformed according to a one-parameter linear-in-log-odds calibration function. The parameter δ_j controls the degree to which the jth participant under- and over-estimates probabilities. We chose a truncated Gaussian prior for this parameter that gives weight to a value of 1 in order to capture sport domains like baseball where the winning percentages are all centered between 0.4 and 0.6 and thus do not need or benefit from a re-calibration function. For the jth participant, δ_j scales π_i according to $\log(\pi_i/(1-\pi_i))$. As a result of this re-calibration, each jth participant perceives the probability judgment for question i to be $\psi_{ij} = \delta_j \log(\pi_i/(1-\pi_i))$.

The calibrated probabilities ψ_{ij} are also affected by the jth person's expertise level. Each person's estimate comes from a Gaussian distribution centered on $exp(\psi_{ij})/1 + exp(\psi_{ij})$ and with a standard deviation of σ_j that corresponds to the jth person's expertise level. We chose an Exponential prior for σ that has a long-tailed distribution allowing for a wide range of values that includes non-experts who are very unknowledgeable about the domain. The behavioral data that we observe from the questionnaire takes the form of the probability estimates p_{ij}, in which each of the participants gave a winning percentage judgment for each of the teams. For each person, all the above processes combine together to generate their probability judgment behavior.

2.3.3 Modeling Results

We implemented the graphical model seen in Figure 2.14 using JAGS. Our modeling results are based on 3 independent chains with 1000 samples after discarding the first 1000 burn-in samples from each chain. The chains were assessed for convergence using the standard $\hat{R}$ statistic (Gelman, 1996).

Model Accuracy

The model inferred the predictions for each ith team's winning percentage π_i. At the end of the season, the true winning percentages became known and we compared our inferred answers to the true answers to see how well the model predicted the probabilities for the entire season by calculating its mean absolute deviation (MAD). To compare our model's

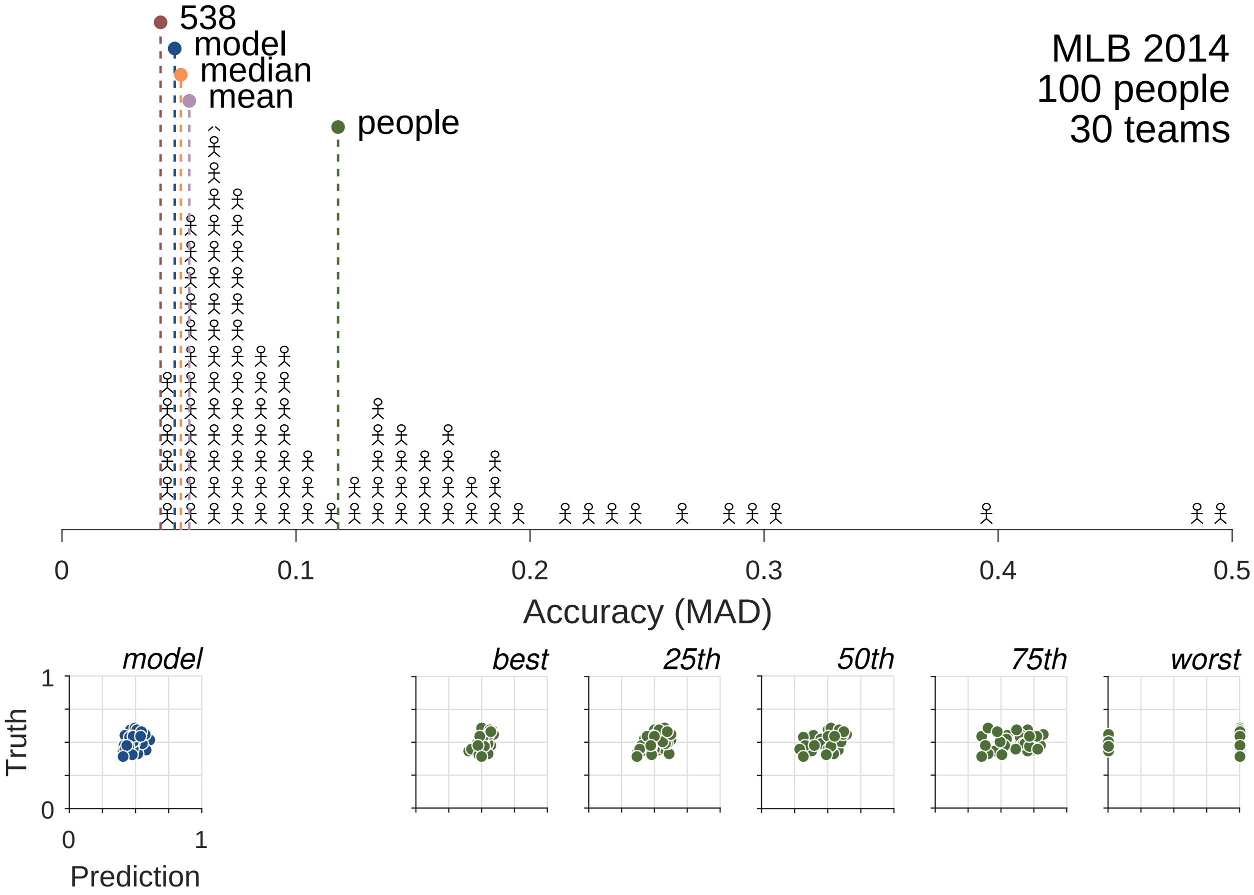


Figure 2.15: The performance of our cognitive model compared to the mean, median, and the fivethirtyeight algorithm in estimating probabilities for the MLB 2014 season. The main panel shows the distribution of individual participant performance as stick figures and the levels of model, mean, median, and the fivethirtyeight algorithm performance as broken lines. The performance of the models and individuals is measured as the mean absolute deviation from true answers.

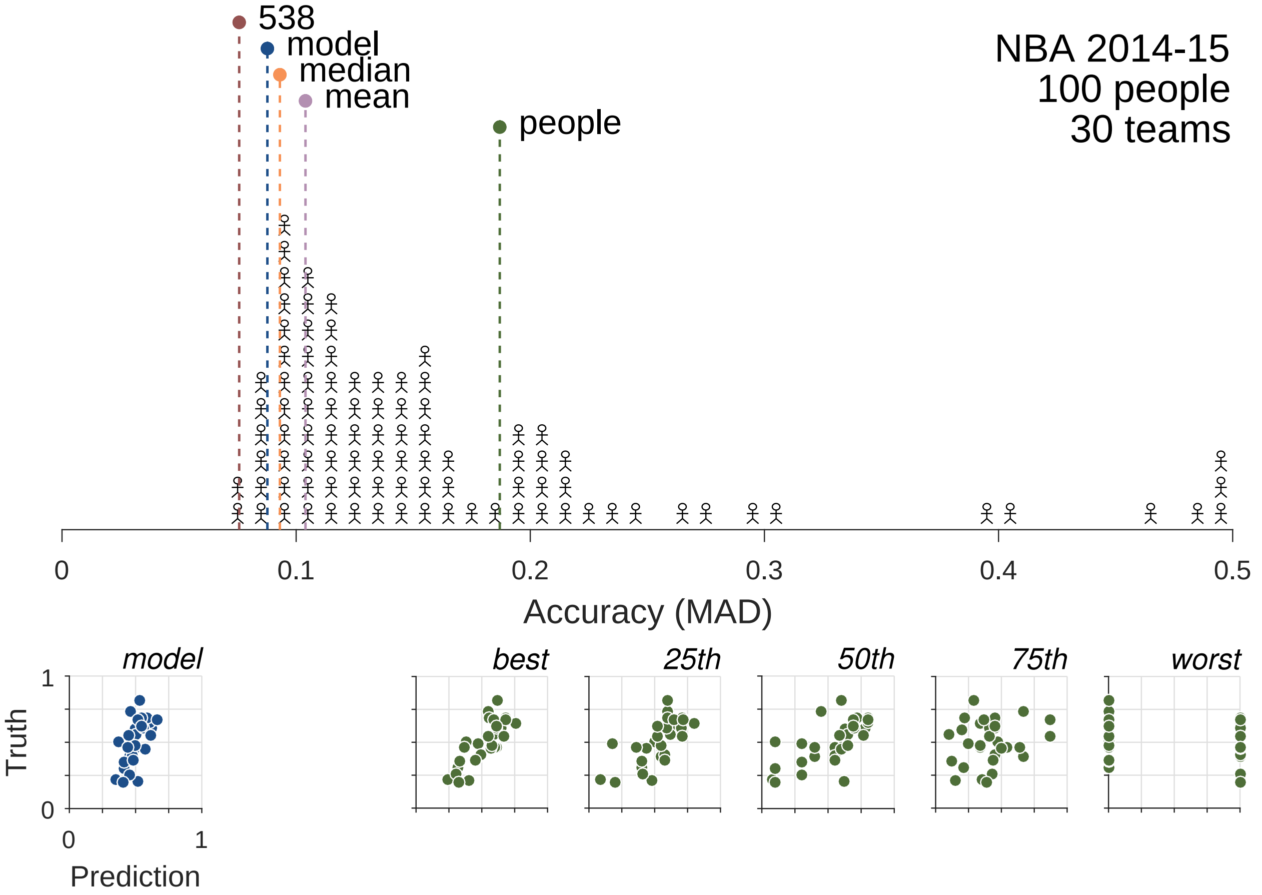


Figure 2.16: The performance of our cognitive model compared to the mean, median, and the fivethirtyeight algorithm in estimating probabilities for the NBA 2014-15 season. The main panel shows the distribution of individual participant performance as stick figures and the levels of model, mean, median, and the fivethirtyeight algorithm performance as broken lines. The performance of the models and individuals is measured as the mean absolute deviation from true answers.

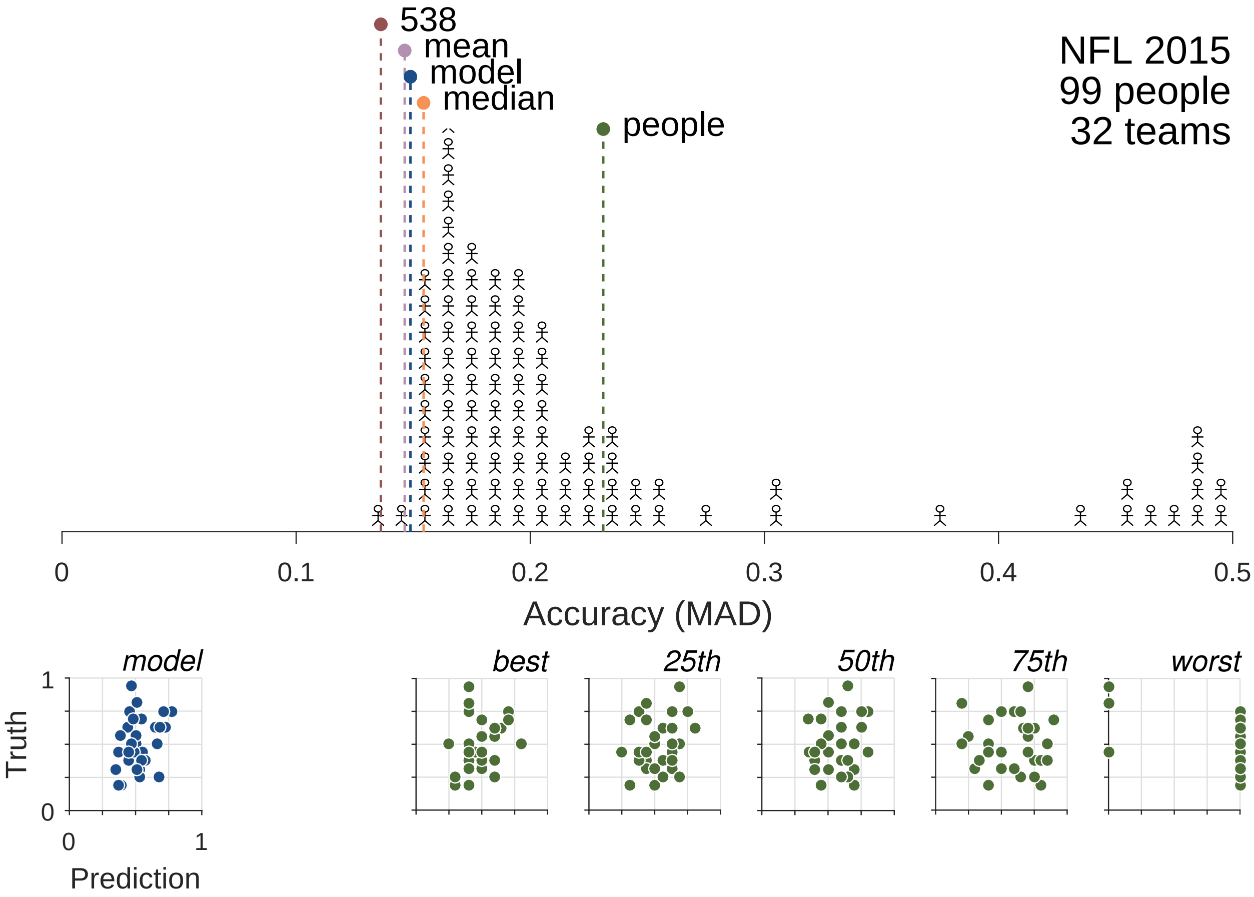


Figure 2.17: The performance of our cognitive model compared to the mean, median, and the fivethirtyeight algorithm in estimating probabilities for the NFL 2015 season. The main panel shows the distribution of individual participant performance as stick figures and the levels of model, mean, median, and the fivethirtyeight algorithm performance as broken lines. The performance of the models and individuals is measured as the mean absolute deviation from true answers.

inferences to traditional statistical aggregation measures, we also calculated the error for the mean and median by taking each of those statistical measures for each team and seeing how well that measure performed as a whole. Additionally, we include metrics from a well-known statistical analysis blog, fivethirtyeight.com that posted predictions for all three of these sports seasons.

We summarize these results in Figures 2.15, 2.16, and 2.17. The main panel shows the distribution of individual participant's accuracies of predicting the correct winning percentages. The vertical dashed lines show the performance of all the participants on average as well as the accuracies of our model, the mean, the median, and the blog fivethirtyeight. For all three sports domains, the best performing method was fivethirtyeight's algorithm, with a MAD of 0.042 for MLB, 0.076 for NBA, and 0.136 for NFL.

For the MLB 2014 season, the second-best performing aggregation method is our model, with a MAD of 0.048, followed by the median with a MAD of 0.051, and the mean with a MAD of 0.054. Since the average of individual accuracies is much worse with a MAD of 0.118, this not only shows that the wisdom of the crowd is an effective method for predicting winning percentages, but that our cognitive model performs very well. For the NBA 2014-15 season, the second-best performing aggregation method is again our model, with a MAD of 0.087. The statistical measures of the median and mean are at 0.093 and 0.104, respectively, while people on average once again have the worst accuracy, with an error of 0.187. For the NFL 2015 season, the second-best performing aggregation method was the mean, with a MAD of 0.146, followed by our model with a MAD of 0.149 and the median with a MAD of 0.154. People's average was the worst with a MAD of 0.231.

The bottom panels in Figures 2.15, 2.16, and 2.17 show scatter plots of the relationship between the winning percentage predictions and the true winning percentages for our model as well as a handful of participants at various percentiles of accuracy. For each of the three sports seasons, we show the best-performing participant, in terms of lowest MAD, the worst-

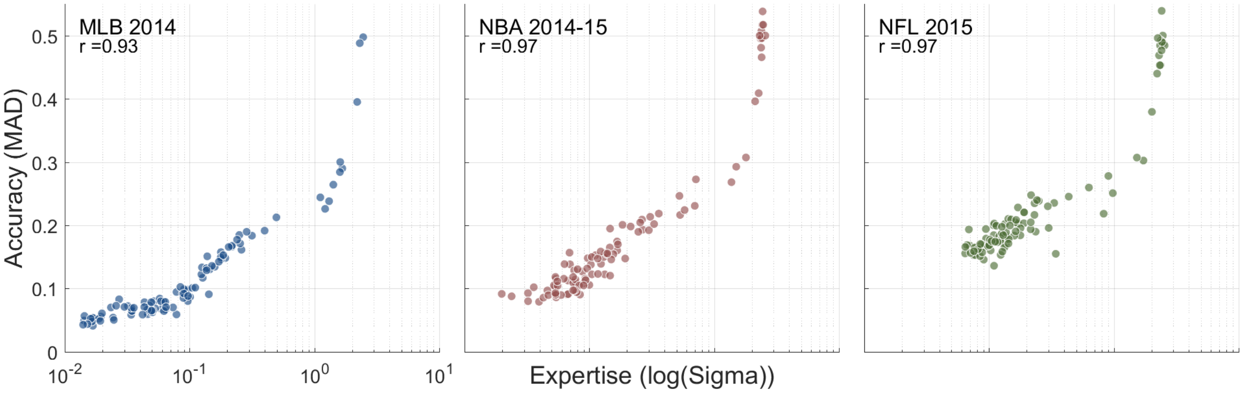


Figure 2.18: Relationship between our model inferences of individual expertise and their accuracy in estimating winning percentages for each of the sports seasons.

performing participant, and participants at the 25th, 50th, and 75th percentile. The panels on the right show the relationship between our model's predictions for the team's winning percentages, or the π values for each team, and the true percentages. If a method perfectly predicted how each team would do in the season, the resulting scatter plot would be a linear one-to-one relationship with an MAD of 0. The worst-performing participants in every sport are those who only gave answers of 0 and 1 instead of a range of possible winning percentages. The other panels show an effect of the calibration curve. Even some the best-performing participants show slight over- and under- estimation of probabilities, best seen in the best-performing participant in the NBA questionnaire. The model's predictions are affected by this as well, since they are more heavily influenced by the ones who are inferred to be experts. This figure shows that despite such large variety in the participants' performance, the model predicts the winning percentages well. In all the cases, the model's error falls between the best-performing and 25th percentile participant.

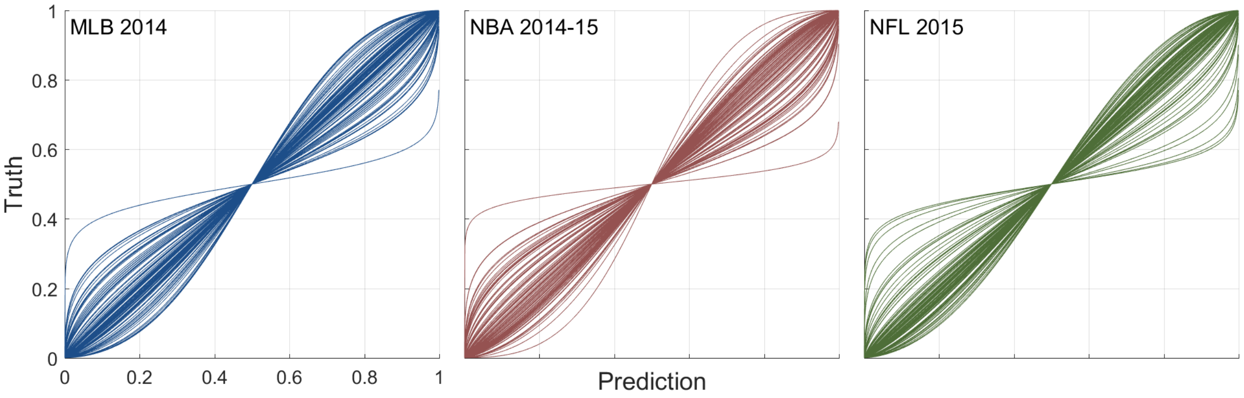


Figure 2.19: Participants' calibration curves from each of the three sports seasons, showing the relationship between the probability estimation predictions and the true probability answers.

Individual Differences

Figure 2.18 shows the relationship between our model's inferred expertise, the parameter σ, for each person and each person's overall performance in terms of MAD. Smaller values of expertise mean that the person is inferred to be more of an expert while smaller values of MAD mean that they predicted the winning percentage results better. There is a strong positive correlation, meaning that people who were inferred to be experts did predict the final results better. This correlation was strongest for the NBA 2014-15 and NFL 2015 seasons, with the correlation coefficient equal to 0.97, followed by 0.93 for the MLB 2014 season.

Figure 2.19 shows the relationship between the probability estimation predictions and the true answers that were known when the sports seasons concluded. Each line is a participant's calibration curve, calculated using the δ that was inferred for each person, controlling the degree of over- and under-estimation. Areas of the graph that are darker mean more participants with that calibration pattern inferred from the data. As evident in Figure 2.19, all three sports have many participants who are prone to mis-calibration of probabilities. The patterns resemble that of Figure 2.12 which showed a framework for how people's probability representations may be different from the latent truth.

2.3.4 Discussion

Our goal was to extend an existing cognitive model of probability estimation aggregation to a truly predictive application for which the answers are not known at the time of data collection. We adjusted prior knowledge of the calibration and expertise parameters in our model to fit the new domains to which the model was applied. Our original model was tested in a domain for which the ground truth was already known, historical events in first-division soccer games, and various world probabilities that could be looked up. The current setting of three sports seasons was a novel application in that the answers could not have been known until after the seasons had ended. We adjusted the calibration parameter δ to allow for a bigger possibility that calibration is not needed, for sports like baseball in which there are no extremely small or large winning percentages. We also adjusted the expertise parameter σ that allows for a bigger number of non-expert participants who may not know anything about sports and should thus be down-weighted in the aggregation. The model then made predictions of the percentages and inferences about participants' cognitive processes *without* knowledge of the answers.

Our model showed a clear "wisdom of the crowd" effect by outperforming the average person's prediction accuracy and it performed as well in this application as standard aggregation measures such as the mean or median. For the MLB and NBA environments, it even outperformed those measures. Even for the NFL environment, the difference in the predictive accuracy (as MAD) of our model and the accuracy of the best performing aggregation method was small. The benefit of a cognitive model for probability aggregation is the additional insight we get into people's estimation behavior.

We are able to identify experts before the true winning percentages are revealed, since the model never sees the answers. People who were inferred by our model to have a higher level of expertise did in fact perform better once the winning percentages were revealed and accuracy was calculated, meaning our model made predictions about the performance of all the individuals. In addition to finding the experts, we also can draw inferences about each person's calibration curve they use to produce their answers. Since many participants in all of the sports domains were inferred to be at least slightly miscalibrated, we have also learned that even the most accurate individuals are prone to biases in estimation, such as the over- and under-estimation of probabilities.

These three sports domains show only a small test of our cognitive model for a predictive setting. Further work can be done for additional testing in a larger variety of sports, or even other domains in which people estimate probabilities. For these applications, more (or less) informative priors can be given to better fit each specific domain. Overall, we expect that our model continues to show a "wisdom of crowds" effect and more importantly reveal valuable insight into cognitive processes and biases behind people's decisions.

2.4 Conclusion

While there is clearly much more that can be done, we think the current results from the above two sections have a clear message. The wisdom of the crowd phenomenon is rooted in what people know, and so theories and models from cognitive psychology should play an important role. Generic statistical approaches to combining estimates or judgments are simple to understand and implement, and often work well, but leave room for improvement. We hope to have demonstrated an approach for achieving this improvement, based on modeling the way in which people produce estimates. The models we developed and applied relied on basic theoretical assumptions about individual differences in knowledge representation and the decision making processes, and performed well relative to the performance of individuals and statistical methods in estimating the ground truth in two different domains, even in a predictive setting. In addition, the models incorporated psychologically meaningful parameters that permitted useful inferences about the expertise of individuals and their calibration of probabilities. We think the dual promise of improved applied performance and deeper psychological insight means cognitive modeling approaches to finding the wisdom in crowds is a promising approach.

Chapter 3

Cognitive Modeling of the Wisdom of the Crowd in Category Learning

3.1 Introduction

Categorization is a fundamental process of cognitive behavior that has been evolved to an efficient capability. Objects and concepts in the world are are assigned labels according to their shared features or functions. Category learning theories have been developed to explain people's classification behavior. These theories vary in their assumptions of how people treat individual instances of a category. Exemplar theories assume that each instance of a category is stored in memory. On the other hand, prototype theories assume that the instances are averaged to form one ideal representation of a category. An alternative approach to categorization is decision bound theory, which assumes that people separate the stimuli by a boundary instead of remembering all of them or constructing a prototypical instance. Another theory involves rule strategies that define a category through a summary of the members of that category. All of these theories have been able to some degree explain human classification behavior and all remain a viable explanation of category learning for various situations. Some have tried to capture benefits of each of these theories in hybrid approaches that involve representation assumptions of two or more of the previously discussed theories. For example, these might combine the prototype approach with the exemplar one or the exemplar approach with some built-in rule-based strategies. Researchers have devised tasks for studying category learning in an experimental setting in order to test these theories. Various types of tasks are developed to exploit and understand differences between categorization theories. Often times, these involve presenting participants in a task with stimuli that vary on two dimensions and asking them to categorize the stimuli into one of two categories.

Computational models quantify and test these assumptions and their predictions of how people learn and treat category structures. One popular theory-based model called the generalized context model (GCM) (Nosofsky, 1984, 1986) quantifies the exemplar-strategy process. Another theory-based model is a decision bound strategy that has been applied to human categorization as a model called general recognition theory (GRT) (Ashby and Townsend, 1986). Other theory-based models, although less popular than the previous two, are prototype models (Reed, 1972; Smith and Minda, 2001, 2002). The ones I will focus on in the rest of the section are the first two: GCM and GRT. These models provide accurate accounts of human categorization behavior but in different scenarios that I will discuss further on. The reason why these models are so useful is that we can use them to predict future behavior. However in order to do that, we need to improve the inferences that these models make about a given individual's categorization strategy. The way a person classifies categories depends heavily on memory capacity, attention to relevant features, and category boundary distinction. As a result of these individual differences, categorization itself is subject to high individual variation. Given a classification task, two different people with fluctuating working memory capacities who attend to different features of the stimuli may end up with two distinct category boundaries separating the category members. Cognitive models of category learning can be naturally extended to incorporate individual differences by fitting each participant with his or her own parameter values.

In this chapter, I discuss the wisdom of the crowd in the context of category learning, using majority decisions. I tackle this challenge in two ways: first empirically, and then using cognitive models. In the first part of the section, I discuss an empirical approach for testing the accuracy of group learning curves produced by aggregating individual categorization decisions for a number of existing category learning data sets. In general, these aggregate learning curves perform as well or better than the learning curves of most individuals. This is perhaps not surprising, given the empirical success of aggregation in other behavioral tasks, including estimation (Herzog and Hertwig, 2009; Vul and Pashler, 2008), problem solving (Yi et al., 2012), ranking and voting (Selker et al., in press; Lee et al., 2014), and competitions (Lee et al., 2011). Most of these tasks, however, involve sets of largely independent decisions, whereas category learning involves sequences of repeated decisions, sometimes with structure at the individual level based on the progress of learning. Thus, the finding of a wisdom of the crowd effect for category learning extends the generality of the empirical effect.

In the second part of the section, I build on the empirical finding using cognitive models. The modeling approach allows the wisdom of the crowd to be extended to the categorization of new stimuli, for which behavioral data do not exist. The key idea is to use models to make predictions of the categorizations an individual would have produced for the new stimuli. The group majority can then be formed across these predictions. I demonstrate this approach in two case studies, involving two different models of categorization, and two different stimulus sets. I then discuss a further extension to the mentioned cognitive models by allowing for individual variation in model use within the same experiment. I conclude by discussing implications of these studies for the field of category learning.

3.2 Empirical Analysis of Wisdom of the Crowd in Category Learning

Imagine that a team of trainee doctors view a set of skin patches, and have to categorize them as being malignant or benign. These doctors receive feedback about their responses and, over time, learn to classify skin patches accurately. Presumably they learn which skin patch dimensions, such as color, size, or shape, are important. In addition, they may learn which levels of these dimensions indicate malignancy. A large skin patch that has a light color and a smooth outline might be benign, whereas a small skin patch that has a dark color and a jagged outline might be malignant. It is likely there will be differences in exactly which patches each doctor sees, or at least the sequence in which they see them. It is also likely that there will be individual differences in how well and how quickly the doctors learn to categorize.

The "wisdom of the crowd" is the phenomenon in which an aggregated group answer to a problem is more accurate than the answer of individuals in the group (Surowiecki, 2004). There are at least two ways an aggregate answer can improve upon an individual answer. One way is *signal amplification*, in which combining answers amplifies the common signal and reduces noise. For example, if a skin patch is malignant, that ground truth provides a common signal that competent doctors will reliably detect, while other doctors will be less consistent in their categorizations. The net result is that the group overall will favor the ground truth of malignancy, even if some individuals believe it to be benign. A second way is *jigsaw completion*, in which different individuals solve different parts of the problem. For example, if there are various types of malignant patches, different doctors may specialize in different types. Relying on the categorizations of the doctors who specialize in each individual patch will maximize the accuracy of the group classification across all patches.

Surowiecki (2004) identifies four requirements for a wise crowd. The first is *diversity*: the individuals need to have a range of different opinions and backgrounds. As the doctor example makes clear, this will often be true of categorization problems, because of individual differences in learning. In general, some people may learn more quickly than others, and some people may achieve eventual levels of categorization accuracy that are higher than other people's. It is also possible that not just the rate and final level of learning will differ, but the nature of the learning itself will differ, with some people learning incrementally and gradually improving their accuracy, and others switching between strategies, leading to sudden changes in accuracy. The second is *decentralization*: the individuals need to draw on different information sources. The doctor example again makes clear that categorization often satisfies this requirement. In general, the doctors will learn from different sets of skin patches, or experience them in a different order. The third is *independence*: the individuals cannot know too much about what others think, so that they provide additional or different information to the group. If doctors are trained in an individual setting, or are otherwise unaware of the categorizations of the other trainees, this requirement will also be met.

Given that categorization satisfies these three requirements, applying the wisdom of the crowd idea hinges on satisfying Surowiecki's fourth requirement. This is *aggregation*: there must be a method for aggregating individual decisions into a group decision. Since categorization decisions are discrete (usually binary), the simplest method of aggregation is to take the majority decision. There is evidence, despite its simplicity, that the majority can lead to accurate and robust decisions, for both low-level perceptual and higher-order cognitive stimuli (Hastie and Kameda, 2005; Sorkin et al., 2001).

3.2.1 Data

To test the accuracy of majority group decisions, we examine 28 existing experimental data sets from a set of previous studies. These data sets were collected with ethical approval

Dataset	n_p	n_s	n_b	n_c	Stimuli
Kruschke (1993)	160	8	8	4	Rectangles
Lewandowsky (2011)	113	8	12	6	Shapes
Navarro et al. (2005)	40	25	8	4	Faces
Zeithamova & Maddox (2006)	170	80	5	4	Gabor patches
Lee & Navarro (2002)	22	9	varied	4	Shapes
Bartlema (2013)	34	8	40	1	Shepard circles
Bartlema et al. (2014)	31	8	40	1	Shepard circles
Smith & Minda (1998) Exp. 1	32	14	7	2	Nonsense words
Smith & Minda (1998) Exp. 2	32	14	10	2	Nonsense words

Table 3.1: Details of the experimental category learning data sets (n_p: number of participants across all conditions of the experiment; n_s: number of stimuli; n_b: number of blocks; n_c: number of conditions)

from the relevant academic institutions. Table 3.1 details the studies, including information about the total number of participants, the number of stimuli, the number of blocks (a set of trials typically presenting each stimulus once), the nature of the stimuli, and the number of experimental conditions. It is the total number of experimental conditions that totals the 28 data sets. These studies were chosen because they were the only ones for which we could find behavioral data at the level of individual participants and individual trials, and the true category membership of each stimulus is known.

As Table 3.1 shows, the data sets vary widely in all of these properties, especially in the nature of the stimuli. The stimuli include rectangular shapes of different sizes (Kruschke, 1993a), shapes varying in size and color (Lewandowsky, 2011), adult faces categorized in terms of gender, hair color, and trust (Navarro et al., 2005), Gabor patches varying in frequency and orientation (Zeithamova and Maddox, 2006), shapes varying in color and form (Lee and Navarro, 2002), Shepard circles of varying size and radial line angle (Bartlema, 2013; Bartlema et al., 2014), and nonsense words (Smith and Minda, 1998).

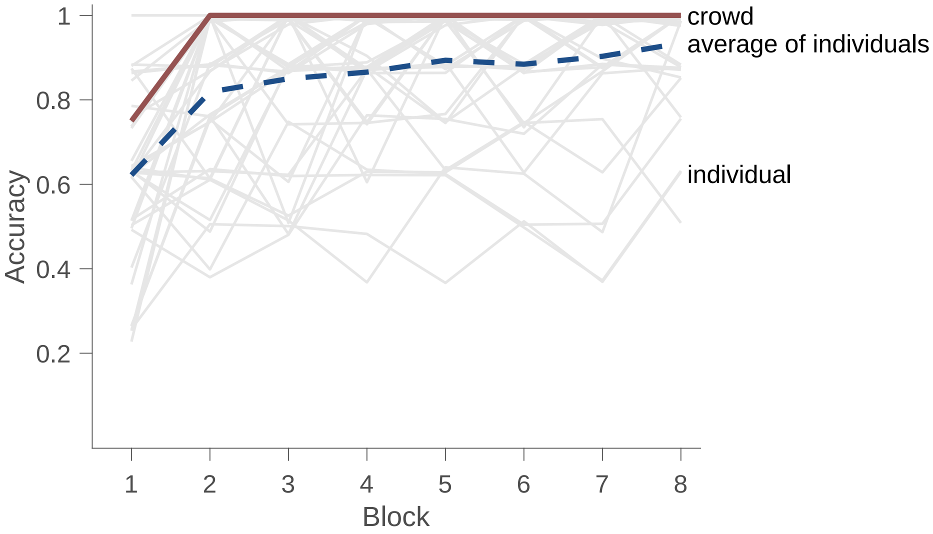


Figure 3.1: Learning curves for one experimental condition from the Kruschke (1993a) data set. The thin gray lines show each individual's proportion of correct answers for each block of the experiment. The dashed blue line shows the average of the individual participant accuracies. The single thick red line shows the categorization accuracy of the aggregated crowd majority decision.

3.2.2 Empirical Analysis

Figure 3.1 shows the results for one data set coming from the Kruschke (1993a) study. The x-axis shows the eight blocks of learning trials, and the y-axis shows categorization accuracy. Because there are two categories, an accuracy of 0.5 corresponds to chance performance. The thin gray lines show the performance of each individual participant, plotting their proportion of correct categorization decisions in each block of the experiment. The average of these individual participant accuracies is shown by the dashed blue line.

Figure 3.1 also shows the performance of the wisdom of the crowd aggregate. The aggregated decisions categorize a stimulus on each trial, just as individual participants did. The difference is that the aggregate decision is based on the majority of the observed participant behaviors for that stimulus. The single thick red line shows the categorization accuracy of these aggregated majority decisions over the course of the experiment. The learning curve for the crowd achieves perfect accuracy as early as the second block of the experiment. This contrasts favorably with individual performance since only a few participants do slightly better in the first block, and clearly is superior to the average performance of people.

Figure 3.2 shows the same analysis for all of the conditions in all of the data sets from Table 3.1. Some experimental conditions are easier to learn, while others are harder. For example, the Bartlema et al. (2014) conditions are difficult because of the perceptual confusability of the stimuli, whereas the fourth Navarro et al. (2005) condition is difficult because the stimuli were randomly assigned to categories. In addition, some experimental conditions show clear evidence of individual differences. For example, in several of the Zeithamova and Maddox (2006) conditions, there appears to be two groups of participants, one learning the category structures and reaching high accuracy, and another failing to learn and remaining at poor accuracy throughout the experiment. Despite this variability, the red lines in Figure 3.2 show that the crowd performs well. For nearly all of the experimental conditions, the crowd outperforms most or all of the individuals, and almost always outperforms the individual average.

3.2.3 Discussion

In the section above, I demonstrated a wisdom of the crowd approach to categorization. The idea is to use the majority categorization decision over a set of individuals as the crowd decision. I showed that this approach leads to accurate crowd decisions for a number of existing category learning datasets, varying widely in the size of the crowd, the difficulty of the category structures, and the nature of the stimuli. The basic result is that taking the majority decision is an effective aggregation method for category learning tasks.

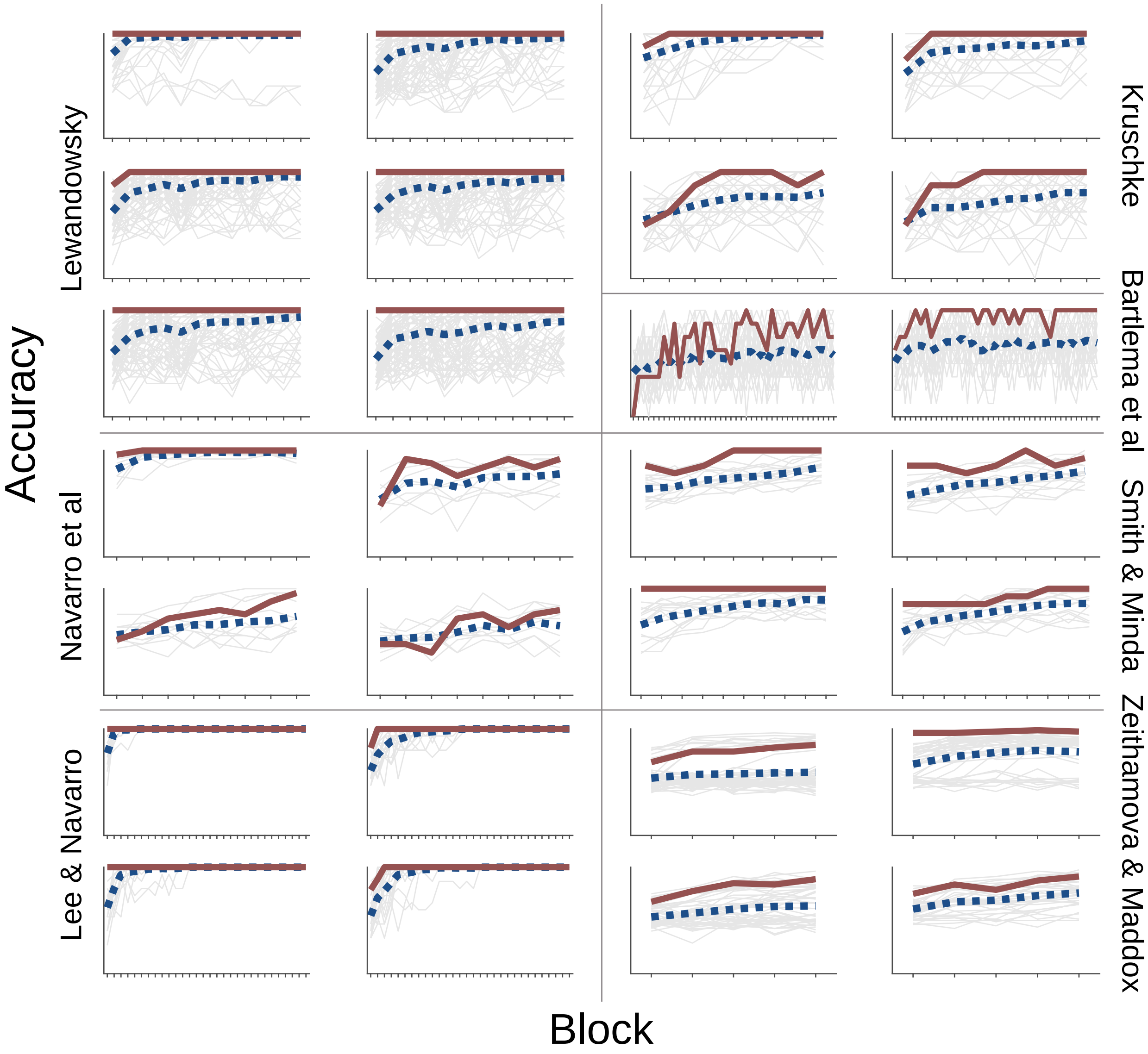


Figure 3.2: Learning curves for 28 category learning experiments from eight data sets. As in Figure 3.1, the gray lines show individual participant accuracy, the dashed blue lines show the average of the individual participant accuracies, and the red line shows the accuracy of the aggregated crowd majority decisions.

3.3 Implementation of Category Learning Cognitive Models

Imagine now that there is a new skin patch that has not been categorized by any of the trainee doctors. In this case, it is not possible to aggregate observed categorization decisions, and so the behavior-based wisdom of the crowd approach does not apply. If it is possible to predict what each doctor would decide, however, these predictions can be aggregated as if they were behavioral decisions.

In this section, we develop a model-based approach for extending wisdom of the crowd categorization to new stimuli. The idea is to infer a cognitive model of each individual's categorization process based on their decisions for stimuli they have seen, and use that model to predict their decisions for new stimuli. We present two examples of this approach, using two different prominent models of categorization, and involving two different types of stimuli. The first uses General Recognition Theory (GRT) (Ashby and Townsend, 1986) and simple perceptual stimuli, while the second uses the Generalized Context Model (GCM) (Nosofsky, 1984, 1986) and face stimuli.

3.3.1 An Application Using General Recognition Theory

General Recognition Theory (GRT), assumes that categorization decisions are made based on decision bounds. For example, in a categorization task in which a person places a stimulus into one of two categories on each trial, GRT assumes decisions are based on a boundary that splits the stimulus space into two response regions. The decision-bound modeling approach is naturally contrasted with exemplar models of categorization, which assume that people remember all instances of a category and keep them in memory for comparison to novel stimuli to make categorization decisions.

An important issue for any model of categorization relates to the possibility of individual differences. Different people may categorize differently, perhaps as a result of different starting knowledge, different training or learning experiences, different learning strategies, or different decision strategies. Many applications of category learning models ignore individual differences, and deal with behavioral data that are aggregated or averaged over people. Other applications apply models at the level of individual participants (Nosofsky, 1986). Most recently, there have been some attempts to extend categorization models to include models of individual differences (Bartlema et al., 2014), using Bayesian methods, but these are restricted to exemplar and prototype models.

For decision-bound models, one important potential source of individual differences relates to the use of unidimensional versus multidimensional boundaries. A working hypothesis in the decision bound literature is that simple category structures that separate stimuli based on a single dimension are amenable to simple explicit rules that can be verbalized, whereas more complicated category structures that require integration across the dimensions need associatively learned boundaries that are more implicit. As a result, one focus of modeling individual differences using GRT is to infer whether people use a simple horizontal or vertical bound that partitions stimuli along one stimulus dimension, or a more general linear (diagonal) decision bound that is sensitive to both dimensions (Ell and Ashby, 2012). This modeling often also considers the possibility of some form of random responding, to identify contaminant participants.

Methodologically, GRT models that incorporate the possibility of individual differences (Ell and Ashby, 2012; Soto et al., 2015) rely on maximum likelihood methods for parameter estimation, and model selection criteria like the Bayesian Information Criterion. While useful, these methods are limited. Maximum likelihood estimation does not allow for the uncertainty in where a person places a decision bound to be inferred, even though there will always be uncertainty remaining after observing their performance on a limited number of trials. Information criteria attempt to correct for the complexity of different possible decision strategies, but do so in an approximate way that equates model complexity with counts of parameters. Using Bayesian methods automatically overcomes both of these limitations.

In this subsection, I demonstrate a novel Bayesian latent mixture approach to modeling individual differences within the GRT framework. I describe our formulation of a model, with six possible categorization strategies, in latent mixture terms to allow for individual differences, and its implementation as a graphical model to allow for fully Bayesian inference. I then examine the inferences this model makes about individual differences in the decision strategies and decision bounds for two experimental conditions in a categorization data set.

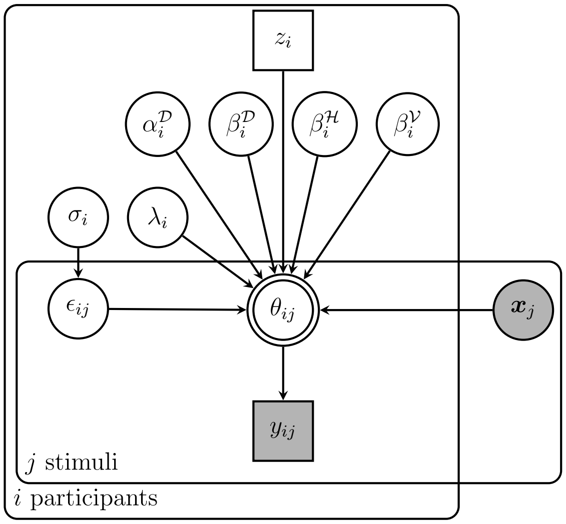
Implementation

Latent mixture models assume that observed data arise from a number of different sources, which combine or mix to produce the overall data. In the case of individual differences in categorization, the different sources correspond to the different categorization strategies used by different people. The latent nature of the mixture means which strategy each individual uses is not known, but rather there are latent parameters assigning people to strategies that need to be inferred. Our model includes six categorization strategies that could be applied to the categorization structures in the experiment. The latent mixture modeling methods we use, however, are general, and could naturally be extended or modified to incorporate different assumptions about individual differences in categorization strategies or types of stimuli. The most obvious categorization strategies to include, in the context of GRT, are vertical boundaries, which are applicable to the unidimensional category structure, and general linear (diagonal) decision boundaries, which are applicable to the information integration structure. We also decided to include a horizontal boundary strategy for completeness. The other three categorization strategies we consider correspond to contaminant models. In the latent mixture approach, with its focus on generative modeling of observed behavior, contaminants are not "removed" by processing the data on the basis of accuracy or other summary criteria, but by modeling the contaminant behavior itself (Zeigenfuse and Lee, 2010). We allow for three different types of contaminant behavior. One corresponds to guessing, in which the participant is equally likely to categorize any stimulus as belonging to Category A as Category B, and the other two assume that all, or almost all, of the stimuli are repeatedly placed in either Category A or Category B.

GRT assumes that decisions are based on a decision boundary that divides the stimulus space into two categories. Each stimulus is represented as a point that defines its location in this space. In this application, the point $x_j = (x_{1j}, x_{2j})$ represents the spatial frequency and spatial orientation of the jth stimulus. GRT assumes that there is variability in the perceptual information associated with each stimulus point on each trial. To account for this, the representation is adjusted to include perceptual noise, so that $x_{pj} = x_j + \epsilon_p$. Categorization decisions are based on which side of a decision bound this point lies.

The decision bound is a discriminant function of the two dimensions that satisfies the implicit line equation $h\left(x_1, x_2\right) = b_1 x_1 + b_2 x_2 + c$, with the three parameters, b_1, b_2, and c. GRT assumes that there is criterial noise ϵ_c added to the discriminant function to account for variations in how the participants remember the bound. It also allows for category bias δ, which can be conceived as shifting the decision bound to favor one category over the other. Putting these assumptions together, the probability a participant will choose category A is given by $\Pr\left(h\left(x_{pj}\right) + \epsilon_c < \delta\right)$.

Figure 3.3 shows the implementation of the GRT as a graphical model. The node y_{ij} is a count of the number of times the ith participant categorized the jth stimulus into category A. The node $\mathbf{x}_j$ is the point that represents the jth stimulus in the stimulus space. The probability θ_{ij} is the probability that the ith participant categorizes the jth stimulus into category A, and is calculated using the cumulative normal distribution $\Phi\left(\cdot\right)$. Following GRT, this categorization probability is determined by the decision bound the participant



$$
\begin{aligned}
\alpha_i^{\mathcal{D}} &\sim \text{Uniform}\big(-\tfrac{\pi}{2}, \tfrac{\pi}{2}\big)\\
\beta_i^{\mathcal{H}}, \beta_i^{\mathcal{V}}, \beta_i^{\mathcal{D}} &\sim \text{Uniform}\big(-\tfrac{1}{2}, \tfrac{1}{2}\big)\\
\sigma_i &\sim \text{Uniform}\big(0, 1\big)\\
\lambda_i &\sim \text{Uniform}\big(0, 10\big)\\
\epsilon_{ij} &\sim \text{Gaussian}\big(0, \tfrac{1}{\sigma_i^2}\big)\\
z_i &\sim \text{Categorical}\big(\tfrac{1}{6}, \ldots, \tfrac{1}{6}\big)\\
\theta_{ij} &= \begin{cases} \frac{1}{2} & \text{if } z_i = 1\\ 0.01 & \text{if } z_i = 2\\ 0.99 & \text{if } z_i = 3\\ \Phi\big(\big[x_{j1} - \beta_i^{\mathcal{V}} + \epsilon_{ij}\big]/\lambda_i\big) & \text{if } z_i = 4\\ \Phi\big(\big[x_{j2} - \beta_i^{\mathcal{H}} + \epsilon_{ij}\big]/\lambda_i\big) & \text{if } z_i = 5\\ \Phi\big(\big[\frac{x_{j2} - \tan \alpha_i^{\mathcal{D}} x_{j1} - \beta_i^{\mathcal{D}}}{\sqrt{\tan^2 \alpha_i^{\mathcal{D}} + 1}} + \epsilon_{ij}\big]/\lambda_i\big) & \text{if } z_i = 6 \end{cases}\\
y_{ij} &\sim \text{Bernoulli}\big(\theta_{ij}\big)
\end{aligned}
$$

Figure 3.3: Graphical model representation of our model for inferring individual differences in categorization using GRT.

uses and the criterial and perceptual noise for the trial on which the jth stimulus was presented. Our model assumes that the criterial and perceptual noise are combined into the value ϵ_{ij} which is drawn from a Gaussian distribution with mean 0 and a participant-specific standard deviation σ_i. Our model also assumes that the category bias δ is equal to 0, since the number of stimuli in each category are equal, and we expect people to be unbiased under these circumstances.

We assume there are, however, individual differences in the decision bounds that people use. In particular, we allow for simple unidimensional categorization strategies corresponding to strictly horizontal or vertical decision bounds, as well as more general diagonal bounds that involve both stimulus dimensions. This is implemented using a latent-mixture approach in which horizontal, vertical, and diagonal bounds are the mixture components. The parameter z_i functions as an indicator variable controlling which type of decision bound the ith participant uses. Because we use a *latent*-mixture approach, the z_i parameter is inferred for each participant. Depending on the type of decision bound, parameters that position that boundary in the stimulus space also need to be inferred. If the ith participant is inferred to use a horizontal bound, it is positioned at a spatial frequency value of $\beta_i^{\mathcal{H}}$. If they use a vertical bound, it is positioned at a spatial orientation value of $\beta_i^{\mathcal{V}}$. If they use a more general diagonal bound, it has a slope of $\alpha_i^{\mathcal{P}}$ and an intercept of $\beta_i^{\mathcal{P}}$. For all of these possibilities, we assume that stimuli are probabilistically categorized according to which side of the decision boundary they lie. Stimuli closer to the boundary are categorized more probabilistically, with some probability they are categorized on the other side of the bound. Stimuli further from the boundary are categorized near deterministically. How quickly probabilistic categorization becomes deterministic is controlled by a participant-specific scale parameter λ_i, as part of a probit-link model of probabilistic responding.

The model in Figure 3.3 also allows for three types of contaminant behavior, motivated by the clear presence of a group of participants exhibiting little learning and responding near chance, as discussed earlier. The first corresponds to the case in which a participant guesses, choosing category A and category B equally often regardless of the stimuli. The second corresponds to the case in which a participant almost always chooses category B regardless of the stimulus. The third corresponds to the case in which a participant almost always chooses category A regardless of the stimulus. Contaminant behaviors can be thought of as alternative response strategies to those coming from GRT, and so are naturally implemented by extending the latent-mixture approach Zeigenfuse and Lee (2010). Thus, overall, the parameter z_i indexes six possibilities for each participant: three possible GRT strategies based on different types of decision bounds, and three possible contamination strategies.

To complete the Bayesian implementation, we set equal prior probabilities on each participant using each of the six possible categorization strategies. We also set uniform prior distributions for the possible range of decision bound locations, and for the noise variability and determinism parameters.

Data

This application is based on two of the experimental conditions reported by Zeithamova and Maddox (2006). These are the two conditions without memory load: the unidimensional condition, which involves 41 participants, and the information-integration condition, which involves 34 participants. In Figure 3.2, these two conditions are the top-right and bottom-right panels in the "Zeithamova & Maddox" section. In both of these conditions, each participant completed five blocks of 80 trials, categorizing Gabor patch stimuli that varied on two dimensions of spatial frequency and spatial orientation. Both conditions gave corrective feedback after every trial, so that the participants could learn to make more accurate categorizations. This application uses data from only the fifth and final block, when participants were the most informed about the category structures.

The two conditions varied in the way the category structures were defined. In the unidimensional condition, stimuli could be accurately categorized solely in terms of their spatial frequencies. The information-integration condition, both spatial frequency and spatial orientation were important for determining the correct categorization. Formally, the stimuli belonging to each category were defined using a multivariate normal distribution over the stimulus space. These distributions allow for the generation of new stimuli from each of the categories for both conditions. Thus Zeithamova and Maddox (2006) provide behavioral decisions for the 80 stimuli in their conditions, but it is possible to generate any number of new stimuli, which participants did not see, but for which their true category membership is known.

Modeling Results

We implemented the graphical model in Figure 3.3 in JAGS Plummer (2003). Our results are based on 6 independent chains with 10,000 samples each after discarding the first 50,000

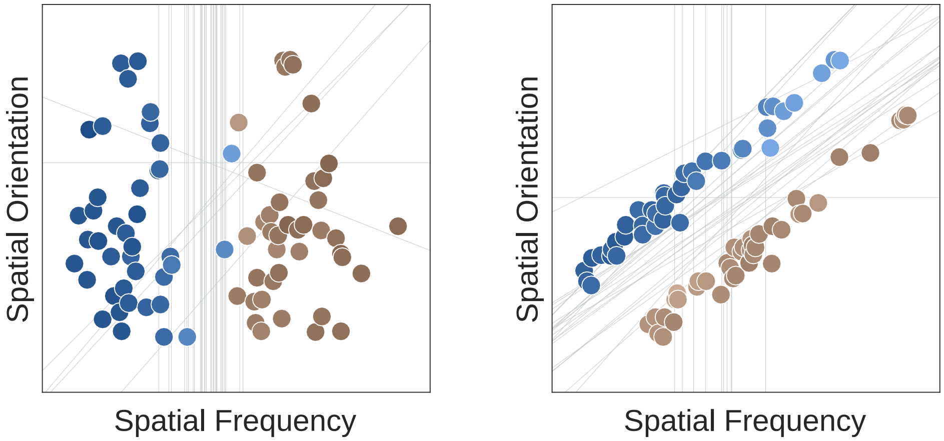


Figure 3.4: Categorization behavior and inferred decision boundaries for the Zeithamova and Maddox (2006) data. The left panel corresponds to the "unidimensional without memory load" condition and the right panel corresponds to the "information-integration without memory load" condition. In each panel, the true category structure is shown by the marker color, and the proportion of correct categorization decisions is shown by shading. The gray lines show the inferred decision bound for each participant.

burn-in samples from each chain and thinning by collecting only every third sample. The chains were assessed for convergence using the standard $\hat{R}$ statistic (Gelman, 1996).

The posterior distribution of the z_i parameter provides the probability that the ith participant is using each of the six possible strategies. We make the simplifying practical assumption that they use the most likely strategy, corresponding to the mode of the posterior distribution. Similarly, for the GRT-based strategies, we assume they use the decision bound given by the posterior mean of the relevant parameters for the horizontal, vertical, and diagonal cases.

These results are summarized in Figure 3.4, which shows the 80 stimuli as points, colored by their true category. The shading of each point corresponds to the proportion of correct categorizations, with darker shades correspond to more accurate decisions. The decision bounds for the participants — 36 in the unidimensional condition, and 33 in the information-integration condition — inferred to be using GRT-based strategies are shown as gray lines. In the unidimensional condition, most participants use a vertical decision bound. However,

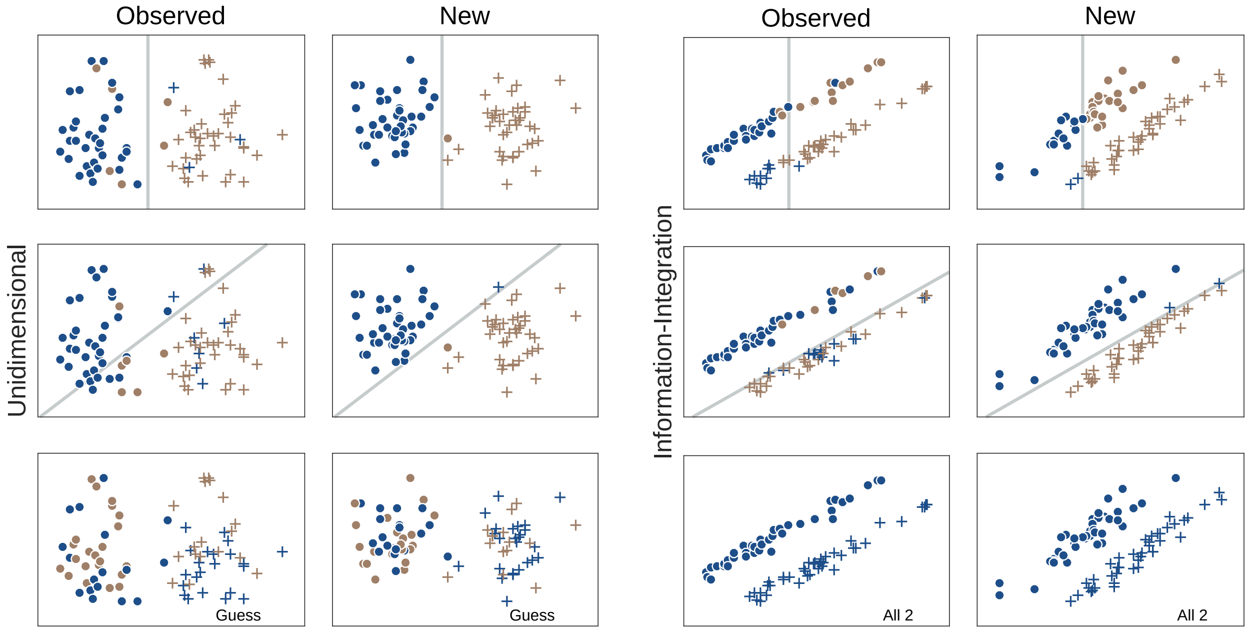


Figure 3.5: Observed and predicted categorization decisions for a subset of the participants from the Zeithamova and Maddox (2006) data set. Panels on the left correspond to the "unidimensional without memory load" condition and panels on the right correspond to the "information-integration without memory load" condition. Rows correspond to individual participants, and columns correspond to observed and predicted behavior. In all of the panels, the marker shape represents the true category. In observed panels, the marker color represents participant behavior in the experiment. In the predicted panels, the marker color represents predicted behavior, based on the inferred categorization strategy. For both conditions, the first two participants are inferred to use the decision bounds shown by the gray line, while the third participant uses either a guessing or repetitive contaminant strategy.

in the information-integration condition, there is a group of participants who use a vertical decision bound and another group who use a diagonal decision bound. In addition to these individual differences in the type of decision bound, there are also individual differences in the location of the bounds themselves. For example, different participants use vertical decision bounds that correspond to different thresholds of spatial frequency.

Figure 3.5 highlights the behavior of three selected participants from both the unidimensional and the information-integration conditions. Each of these participants corresponds to a pair of panels. The "observed" panels show the presented stimuli, with color corresponding to the categorization decision and the marker shape corresponding to the true category. The inferred boundary for the participant is shown by the gray line. This single boundary represents each participant's most likely boundary strategy, either vertical, horizontal, or diagonal, as well as the most likely location of that boundary in the stimulus space. The GRT is able to describe observed behavior to the extent that the decision bound separates the stimuli (i.e., that different colors lie on different sides of the bound). The selected participants vary as to whether they use a vertical bound, a diagonal bound, or one of the contaminant strategies.

The "predicted" panels in Figure 3.5 show how the inferred strategies are applied to make predictions about how each participant would categorize the new set of stimuli. For GRT-based strategies, new stimuli are simply categorized according to which side of the bound they lie. Otherwise, the stimuli are categorized according to the inferred contaminant strategy. It is these predictions that allow us to apply the wisdom of the crowd method to new stimuli.

Wisdom of the Crowd Results

Figures 3.6 and 3.7 summarize our wisdom of the crowd analysis for the unidimensional and information-integration conditions, respectively. For the observed stimuli, the gray bars show

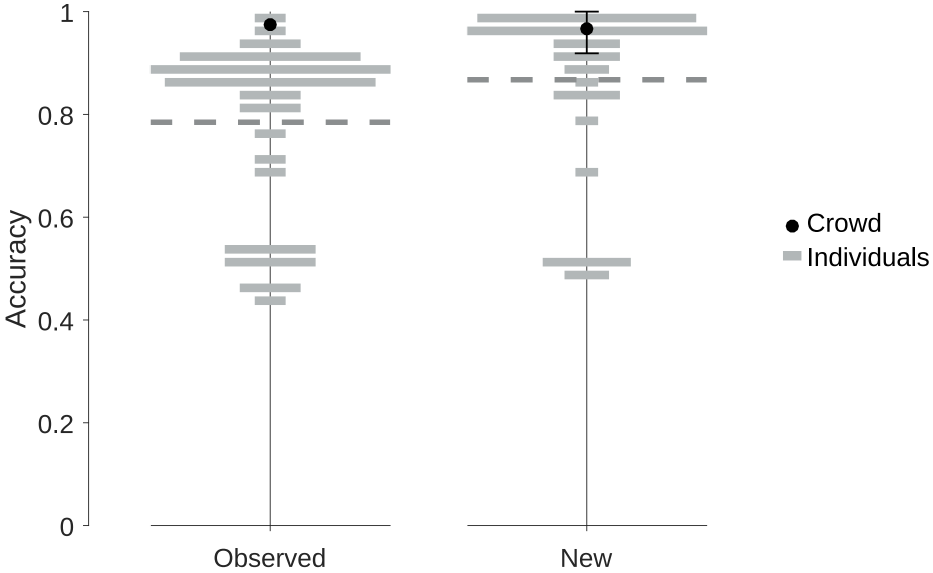


Figure 3.6: Individual and crowd accuracy for the Zeithamova and Maddox (2006) "unidimensional without memory load" condition. For both observed and new stimuli, the gray bars show the distribution of individual accuracy, the dashed line shows the average of these individual accuracies, and the black dot shows the accuracy of the majority crowd decision.

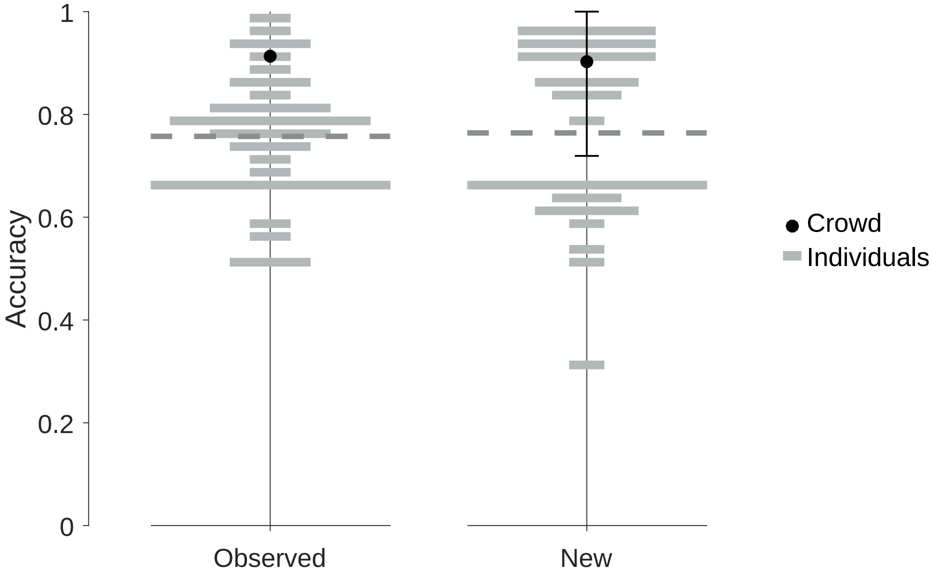


Figure 3.7: Individual and crowd accuracy for the Zeithamova and Maddox (2006) "information-integration without memory load" condition. For both observed and new stimuli, the gray bars show the distribution of individual accuracy, the dashed line shows the average of these individual accuracies, and the black dot shows the accuracy of the majority crowd decision.

the distribution of the categorization accuracy across participants. This is based on their behavioral responses in the experiment. The broken line shows the average of individual accuracy. The black dot shows the accuracy of the majority of these observed categorization decisions.

The right-hand panels of Figures 3.6 and 3.7 involve 1000 sets of newly generated stimuli. For these new sets, the gray bars show the distribution of accuracy for the *predicted* categorization decisions across participants. The broken line is again the average accuracy. The black dot is the average accuracy of the majority predicted categorization across all of the new sets, and the error bar shows its 95-percentile range.

The observed results mirror those presented in Figure 3.2, showing that the majority decision is generally very accurate compared to individual performance. The similarly good performance for the new stimuli shows the effectiveness of the model-based approach. The majority of the predicted decisions, where the predictions are generated by models of individual categorization behavior, is able to categorize accurately stimuli that have never been observed.

3.3.2 An Application Using the Generalized Context Model

The Generalized Context Model (GCM) is an exemplar-based model of categorization that uses selective attention and similarity comparison processes to categorize stimuli. There are several variants of the GCM designed to accommodate specific categorization situations. These situations include some stimuli being presented more frequently than others, category assignment being inherently probabilistic, or the use of stimuli for which perceptual learning is possible (Nosofsky, 1992). The GCM represents stimuli as points in a psychological space that are compared to exemplars in the same space, or instances of other stimuli, stored in memory. Categorization decisions in the GCM are made probabilistically by comparing the

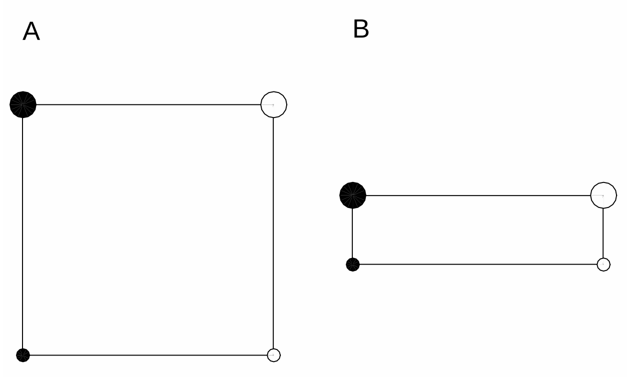


Figure 3.8: Task stimuli varying on size and color(A). Selective attention to only the color dimension (B).

presented stimulus to an exemplar via a similarity function that is influenced by attention and bias. The GCM is an extension of Luce's similarity choice model for stimulus identification (Luce, 1963). Like the GRT, the GCM incorporates individual differences, but instead of people differing in the construction of a boundary, they differ in their selective attention, strategy use, and bias.

Since attention is a process that is vital for many cognitive functions such as memory and decision making, the GCM accounts for the amount of attention that people dedicate to a categorization task. Selective attention to one dimension over the other influences people's categorization decisions. In Figure 3.8A, there are four stimuli that vary along two binary dimensions of color and size. Each side of the square corresponds to a value on one of these dimensions. If a participant selectively paid attention only to the color dimension, the other dimension would be weighted less. In Figure 3.8B, the stimuli of the same size would look more similar to each other than they would in Figure 3.8A because their psychological distance would be smaller. The difference in the shapes' color would be emphasized but the differences in size would be down-weighted. However, if a different participant is attending only to the size dimension, for him or her it would be the color dimension that is down-weighted. Depending on the task, this may help or hinder performance. A model would need to be applied at the individual level in order to provide a good fit to both of these

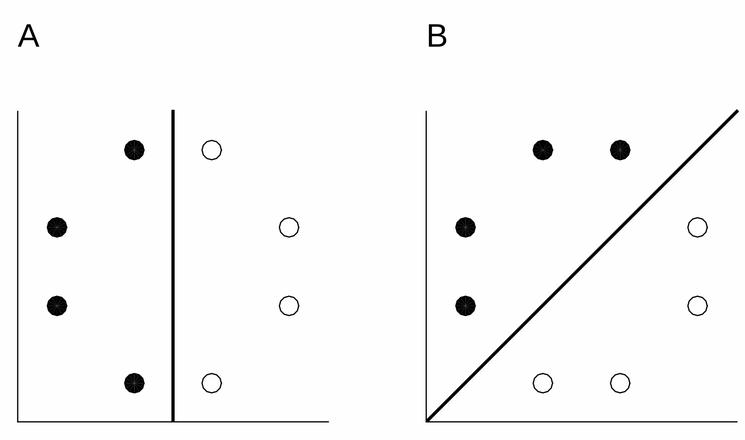


Figure 3.9: A vertical (A) and diagonal (B) decision boundary splitting the stimulus space that has half of the stimuli belonging to category 1 (black circles) and half to category 2 (white circles).

participants. This type of attention process is particularly interesting for categorization task structures that have separable dimensions, since it would be easy for people to attend to one dimension and ignore the other dimension. For integral dimensions, it is very difficult (or even impossible) to only attend to one dimension. Thus it would be beneficial to examine other attention processes susceptible to individual variation.

Another source of individual differences is variation in the position of the category boundaries. If the stimulus space can be separated into two or more dimensions, there can be a wide variability in where the category bounds are placed on this space. As described previously in the paper, rule-based strategies (unidimensional and easily verbalizable) are used in simple category structures that separate stimuli across a single dimension. Multidimensional rules (which are more implicit) are used in more complicated category structures that require integrating information across different dimensions. For various categorization tasks, some participants might be using a unidimensional boundary while others might be using a multidimensional one. The former boundary can be visualized as a horizontal or vertical (Figure 3.9A) bound that partitions the stimulus space. The latter boundary is visualized as a diagonal bound (Figure 3.9B) that is sensitive to the different dimensions.

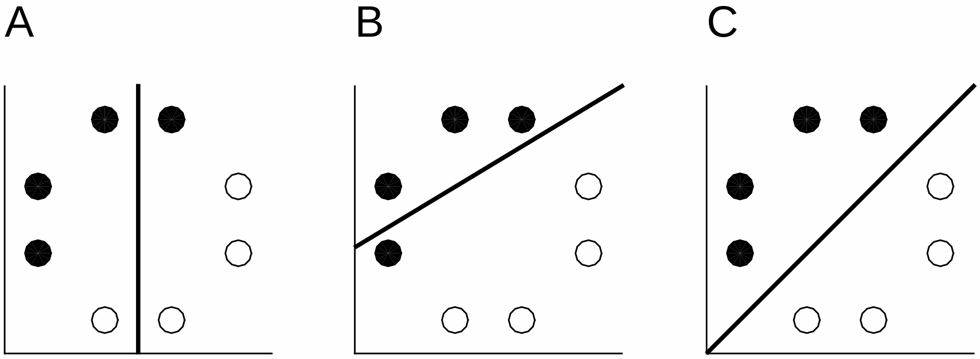


Figure 3.10: An example representation of a participant's decision bound strategy changing over the course of the experiment. The correct representation of the category structure is shown through the stimuli color (i.e the black circles belong to category 1 and the white circles to category 2). In the first trial block (A), he or she is using a suboptimal vertical bound. In the second trial block (B), the bound is adjusted but still slightly suboptimal. In the third block (C), the participant reaches the optimal decision bound to accurately categorize the stimuli.

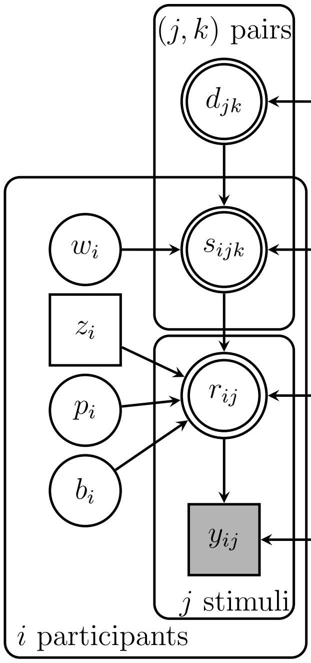
The decision bounds that each person is using to categorize the stimuli might not be stable over the course of the experiment. For example, while the participant is learning an integral-dimension category structure, he or she might start out with using a unidimensional decision bound in the first few blocks. Once he or she realizes that the accuracy is low (if they are given feedback), the participant might adjust the decision bound to a multidimensional one. A schematic of this is presented in figure 3.10. Another participant might be using a horizontal decision bound initially but finally end up at the same diagonal decision bound as the first participant. Although their final accuracy might be comparable, the strategy by which they got there is different.

An aspect of responding that has been incorporated into certain category learning models is the concept of bias. This is a simple concept that examines to what extent each individual in a task is inclined to respond with the same answer throughout all the trials. For example, a person might be less likely to push a button that is designated to response "Category A" than they are to push one designated to response "Category B". The behavior would show a much higher rate of responding one answer than the other. This is vital to take into consideration when also trying to account for contaminants in the data. There is a possibility of a participant in a task simply pushing one button for the entirety of the experiment instead of completing the task. His or her bias toward that answer would be very high and it would be beneficial for a model to capture that type of contaminant behavior.

In this subsection, I describe a Bayesian approach to modeling individual differences within the GCM framework. I describe our formulation of a latent-mixture version of the GCM, with six possible categorization strategies, in latent mixture terms to allow for individual differences, and its implementation as a graphical model to allow for fully Bayesian inference. I then examine the inferences this model makes about individual differences in the decision strategies and decision bounds for an experimental condition in a categorization data set.

Implementation

The GCM has been previously implemented in a Bayesian framework (Vanpaemel, 2009; Lee and Wagenmakers, 2013) but has not been represented as a latent-mixture model for modeling individual differences. Formally, the ith stimulus is represented as a two-dimensional coordinate location $\mathbf{x}_i = (x_{i1}, x_{i2})$. The attention-weighted distance between the ith and jth stimuli is then $d_{ij} = w\,|x_{i1} - x_{j1}| + (1-w)\,|x_{i2} - x_{j2}|$, where w is a parameter controlling how much attention is given to the first dimension. This means that a dimension receiving more attention will be more influential in determining distances than the one receiving less attention. We assume that there may be individual differences in attention, and so there are individual-level w parameters. The similarity between these stimuli is $s_{ij} = \exp\left(-cd_{ij}\right)$, where c is a parameter controlling the generalization gradient. Because we assume individual differences perceptual learning for the face stimuli are unlikely, the same generalization parameter is used for all participants. We do, however, allow c to vary over blocks, allowing for the possibility the degree of generalization is adapted to the learned category structures.



$$\begin{aligned}
w_i &\sim \begin{cases} \text{Uniform(0,1)} & \text{if } z_i = 1 \\ 0 & \text{if } z_i = 2 \\ 1 & \text{if } z_i = 3 \end{cases} \\
c &\sim \text{Gaussian}(1,1)_{T(0,1)} \\
p_i &\sim \text{Uniform}(0,1) \\
b_i &\sim \text{Gaussian}(0.5,0.2)_{T(0,1)} \\
d^m_{jk} &= |x_{jm} - x_{km}| \\
s_{ijk} &= \exp\left\{-c(w_i d^1_{jk} + (1-w_i) d^2_{jk})\right\} \\
z_i &\sim \text{Categorical}\left(\tfrac{1}{6}, \ldots, \tfrac{1}{6}\right) \\
r_{ij} &= \begin{cases} \frac{b_i \sum_j a_j s_{ijk}}{b_i \sum_j (a_j s_{ijk}) + (1-b_i) \sum_j (1-a_j) s_{ijk}} & \text{if } z_i = 1,2,3 \\ p_i & \text{if } z_i = 4 \\ 0.99 & \text{if } z_i = 5 \\ 0.01 & \text{if } z_i = 6 \end{cases} \\
y_{ij} &\sim \text{Binomial}(r_{ij}, n)
\end{aligned}$$

Figure 3.11: Graphical model representation of our model for inferring individual differences in categorization using the GCM.

The similarity of the ith stimulus to category A is then the sum of the similarities to all the stimuli in the category: $s_{iA} = \sum_{j \in A} s_{ij}$. Finally, the probability of a category response placing the ith stimulus in category A is $p_{iA} = bs_{iA} / \left(bs_{iA} + (1-b)\, s_{iB}\right)$, where b is a parameter controlling the response bias to category A. Because the categories are fixed, we do not include a response-determinism parameter in the category response model. We do, however, allow for individual bias, consistent with the assumption that there may be individual differences in the way participants learn the unequal category sizes.

Figure 3.11 shows the implementation of the GCM as a graphical model. Unlike the GRT application, we consider every block in the category learning experiment. Since the GCM does not model learning, we did this by applying it cumulatively over the sequence of blocks. In the graphical model, the y_{ij} node counts the number of times the ith participant categorizes the jth face into category A. The cumulative approach means that this count includes the current block as well as all previous blocks, and n counts how many times it has been presented over these blocks. Following the GCM, the category A response probability r_{ij} is determined from the similarities s_{ij} for the kth participant, which in turn are determined from the distances d_{jk} between the stimulus representations $\mathbf{x}$. The response probabilities depend upon individual bias b_i, and the similarities depend upon the generalization gradient c and individual attention weights w_i.

We assume there are individual differences in the attention weights that people use, and give theoretical weight to the use of simple attention strategies that focus on just one stimulus dimension. We implement this using a latent-mixture approach in which the mixture components are the attention weight w_i values of 0, 1, or drawn from a uniform distribution. An attention weight of 0 corresponds to a person attending only to dimension 2 in Figure 3.12. An attention weight of 1 corresponds to a person attending only to dimension 1 in Figure 3.12. A person inferred to be using an attention weight drawn from a uniform distribution devotes attention to both dimensions, but possibly not equally. Similar to our GRT latent-mixture model, the z_i parameter functions as an indicator variable controlling which attention weight value the ith participant uses.

To complete the Bayesian implementation, we set a prior on the generalization gradient consistent with the distribution of distances in the MDS representation, and a prior for bias that corresponds to expecting any deviation from unbiased responding to be smaller rather than larger. We also set equal prior probabilities on each of the six possible categorization strategies.

Data

This application is based on one of the four conditions in the category learning experiment reported by Navarro et al. (2005). This experiment involved a set of 25 faces. The four conditions differed in the way these faces were assigned to two categories. We consider

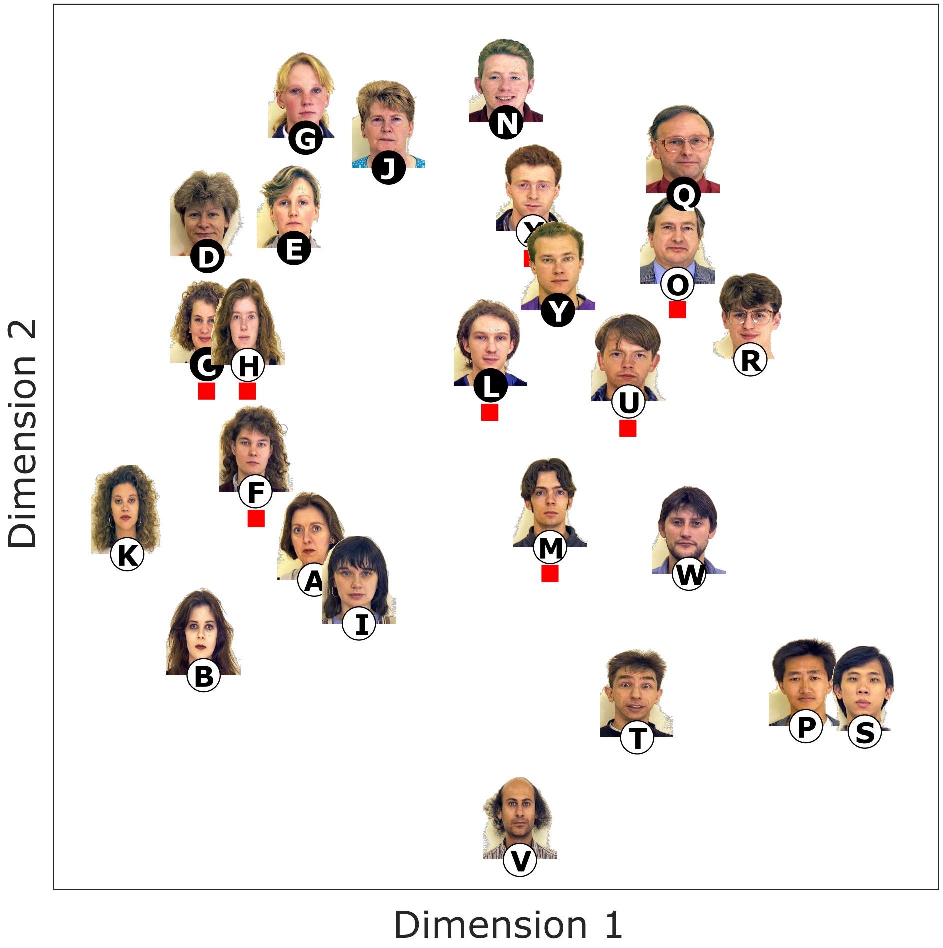


Figure 3.12: Dimensional representation of the face stimuli from the Navarro et al. (2005) data set. The true category structure is shown by filled and unfilled alphabetic labels. The red squares underneath the labels indicate the eight faces that were removed.

only the category structure that divided the faces in terms of hair color. In Figure 3.2, this condition is the top right panel in the "Navarro et al." section. In this condition, 10 participants completed eight testing blocks in which each stimulus was presented once with corrective feedback.

Figure 3.12 shows each of the faces, labeled A–Y, in terms of their representation in a two-dimensional stimulus space. The space was derived by using the individual-differences multidimensional scaling method presented by Okada and Lee (2016), based on previously collected similarity data involving 14 participants rating each pair of faces on a 5-point scale.

A key feature of this multidimensional scaling method is that it derives stimulus spaces with psychologically interpretable dimensions. The dimensions in Figure 3.12 can be interpreted as corresponding to gender and hair color.

The category structure for the hair color condition is indicated in Figure 3.12 by the black and white coloring of the stimulus labels. Unlike the Zeithamova and Maddox (2006) experiment, there is no rule for generating new stimuli with known category assignments. Accordingly, we removed eight faces from the Navarro et al. (2005) data set. The removed faces are highlighted in Figure 3.12 by red squares beneath the stimulus labels. We treat these faces as if they were new stimuli, never seen by the participants.

Modeling Results

We again implemented the graphical model in JAGS. Our results are based on 3 independent chains with 1000 samples each after discarding the first 5000 burn-in samples from each chain. The chains were again assessed for convergence using the standard $\hat{R}$ statistic.

The top panel of Figure 3.13 shows the most likely model for each participant in each block. It is clear that some participants change their attention weights over time, and that there are individual differences in these patterns of change. The most common attentional strategy is to attend just to the second dimension. The first two participants always attend to the second stimulus dimension, and the next two participants do the same from the second block onwards. Other participants show guessing contaminant behavior on many of the blocks. It is relatively rare for participants to attend to both dimensions. In general, the patterns of change are interpretable, such as participant 7 who initially attends to the first dimension, then distributes their attention for a few blocks, and finishes by guessing for the remainder of the experiment.

The bottom-left panel of Figure 3.13 shows the inferred similarity gradients, over the dis-

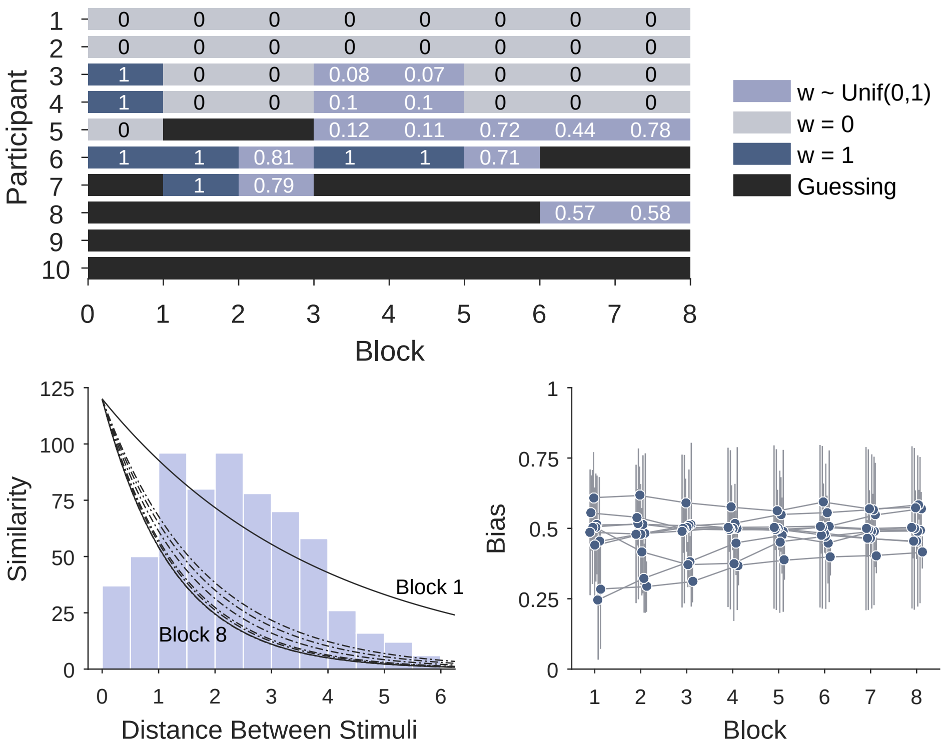


Figure 3.13: Most likely model for all ten participants for all eight blocks shown via color-coding (top panel). Inferred similarity gradients over all stimuli distances in the MDS representation (bottom-left panel). Inferred bias parameter for all ten participants over the eight blocks (bottom-right panel).

tances in the MDS representation of the faces, based on the posterior of the c parameter. The histogram shows the distribution of distances between all pairs of faces. The eight gradients shown correspond to the eight blocks, and the first and last blocks are labeled. The gradient narrows over the course of the experiment, consistent with some form of adaptation. Comparing the gradients to the distribution of distances shows that in the first block, there is broad generalization from one face to all other faces, but in later blocks, what is known about one face generalizes only to relatively nearby faces. This is consistent with the principle of semi-distributed representation (Kruschke, 1993b).

The bottom-right panel shows the inferred pattern of change in bias for each of the 10 participants over the course of the experiment, with error bars representing 95% credible intervals. A few participants show some small initial bias, but the general result is that most participants on most blocks do not favor one category response over the other.

Wisdom of the Crowd Results

We used the GCM, with inferred individual differences in attention and bias, to make categorization predictions for each of the withheld faces from Figure 3.12. As before, the prediction is the most likely category response and the crowd categorization decision is the majority of the individual-participant predictions. For each individual, we used the most likely strategy that the model inferred they were using to generate their categorization decision for the withheld stimuli. When that most likely strategy involved the GCM, we used attention and bias parameters corresponding to the inferred posterior mean for the individual. We then took the modal predicted categorization decision from the non-contaminant participants for each withheld face for each block of the experiment. This final step generated the crowd categorization decision.

Figure 3.14 summarizes the wisdom of the crowd analysis. It shows the average categoriza-

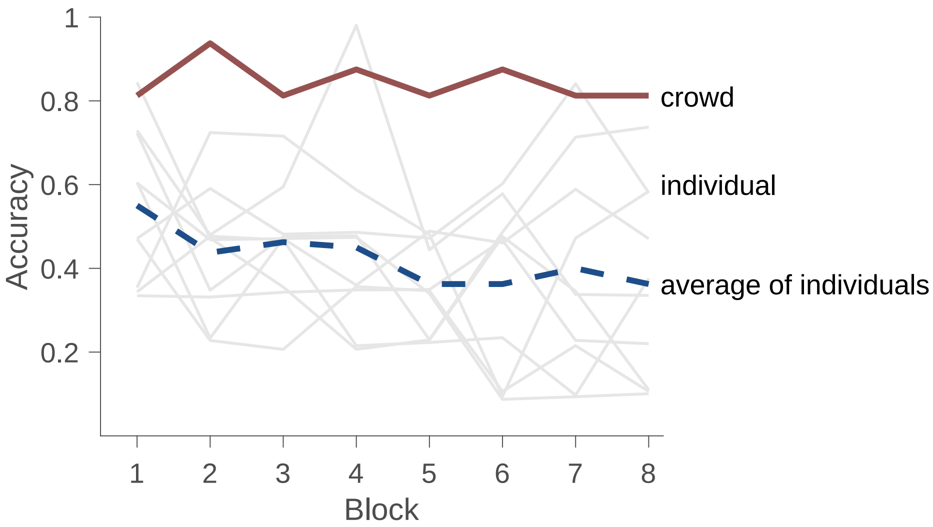


Figure 3.14: Learning curves for the removed faces from the Navarro et al. (2005) data set. The thin gray lines show each individual's proportion of correct answers for each block of the experiment. The dashed blue line shows the average of the individual participant accuracies. The single thick red line shows the categorization accuracy of the aggregated crowd majority decision.

tion accuracy for the withheld faces for the individual participants (i.e. the categorization decisions actually made by the participants before we removed them for modeling purposes), the average of these individual accuracies, and the crowd category decisions. It is clear that from the first block, the crowd is more accurate than any individual and maintains this superiority over the subsequent blocks. The crowd is always more accurate than the average of the individuals. The performance of the aggregate decision is especially impressive, given the difficult of the withheld faces in terms of the category structure, as evidenced by the average of individual accuracy decreasing even with feedback.

3.3.3 Discussion

In this section, I developed a model-based extension of the wisdom of the crowd idea, using categorization models that allow for individual differences in categorization behavior. I showed that individual-level models can be inferred from available categorization decisions and then used to predict how that individual would categorize an unseen stimulus. Our results show that the majority of these predictive decisions continues to produce relatively accurate crowd decisions.

The two case studies we presented highlight the potential generality of the approach. One involved General Recognition Theory and decision-bound categorization, while the other involved the Generalized Context Model and similarity-based exemplar categorization. One involved low-level perceptual Gabor patch stimuli, while the other involved more complicated and holistic face stimuli. One focused on individual differences in the form of different decision strategies, such as horizontal, vertical, and diagonal decision bounds, while the other focused on individual differences in the form of selective attention to different stimulus dimensions. The basic approach simply needs a predicted decision for each individual for a new stimulus, and any model of categorization decisions and individual differences is potentially applicable.

The particular versions of the GRT and GCM we used worked effectively, but we do not claim they are the best possible models. In both case studies, we made a number of modeling decisions, about the inclusion or exclusion of parameters in the GRT and GCM, about the existence of contaminant sub-groups, and so on. These decisions usually had some basis in theory or the specific nature of the category learning task. For example, we did not allow the generalization parameter in the GCM to vary across people, because that would imply some individual variation in perceptual learning over the course of the experiment, which we think is unlikely for the face stimuli. Similarly, our GRT model did not include a category bias parameter, because the number of Gabor stimuli in each category was equal, consistent with what we would expect participants to assume, but we did include such a parameter in the GCM model, because the number of face stimuli in each category was unequal, and we expect individual differences in participants learning this imbalance.

Despite these sorts of justifications, however, it would be possible to explore a large number of alternative GRT and GCM models by combinatorially varying the assumptions we made. This would be interesting theoretically, to test which assumptions are key to good performance, and useful practically, to optimize performance. We noted some interesting possibilities in constructing our case studies, but did not attempt a systematic investigation. For example, in the GCM analysis, we observed that crowd performance was significantly worse before we included the contaminant behavior mixture components. Without allowing for these individual differences, the crowd performance did not go above an accuracy of 75%. Removing contaminants reduces the number of decisions contributing to the majority, but evidently this deficit is more than compensated by identifying those participants who are learning the category structure. In this case, the additional theoretical complicated of including contaminant behavior was worthwhile. It might also be that sometimes a simpler model is a better account of people's behavior, and improves performance. For example, even though the possibility of individual differences in category bias for the GCM case study was well motivated, the inferences in Figure 3.13 suggest that assuming unbiased responding for all participants might describe the data well, and could potentially lead to better crowd performance. Exploring these sort of possibilities is an interesting direction for future research. It will be challenging territory to navigate, because of possible tensions between modeling assumptions that follow from established theory, those that are required to describe the current behavioral data, and those that best achieve the applied goal of crowd accuracy. Ultimately, we need to understand potentially complicated relationships between the quality of a cognitive model of individual categorization behavior, the quality of a model of individual differences in that behavior, and the quality of the crowd performance it underpins.

3.4 Individual Variation in Model Use

The Generalized Context Model assumes people store exemplars of each category in memory, attend to the relevant dimensions of the stimuli, and categorize a stimulus using similarity-based generalization from these exemplars. In contrast, General Recognition Theory assumes that people use a decision bound to partition the stimuli into discrete categories. Categorization seems likely to depend heavily on psychological components and processes such as memory capacity, attentional control, decision-making biases, and so on, all of which may vary across people. Accordingly, it seems reasonable to expect meaningful individual differences in categorization, and this expectation is supported by model-based and empirical evidence (Bartlema et al., 2014; Soto et al., 2015). Previous work has studied different types of strategies within models like the GCM and GRT. For example, previous GRT modeling has emphasized the possibility that different people might use different decision bounds, including special cases like unidimensional horizontal or vertical bounds (Ashby and Gott, 1988; Maddox and Ashby, 1993).

In this section, I present an approach to inferring the general models and specific strategies people use in categorization tasks. I do this by allowing for individual differences between the GCM and GRT models, and for individual differences in specific strategies, like unidimensional bounds, possible within each model. We develop a latent-mixture modeling approach that infers the model and strategy each person is using. Building on previous work in which both the GCM and GRT have been implemented as Bayesian graphical models (Lee and Wagenmakers, 2013; Danileiko et al., 2015), we implement our latent-mixture approach also as a graphical model, allowing for fully Bayesian inference. We apply our model to four existing categorization experiments–all involving stimuli that can be represented in terms of two underlying psychological dimensions–but with various types of stimuli and category structures. We find evidence for large individual differences both between and within models. We finish by discussing the implication of our model and results for future research in

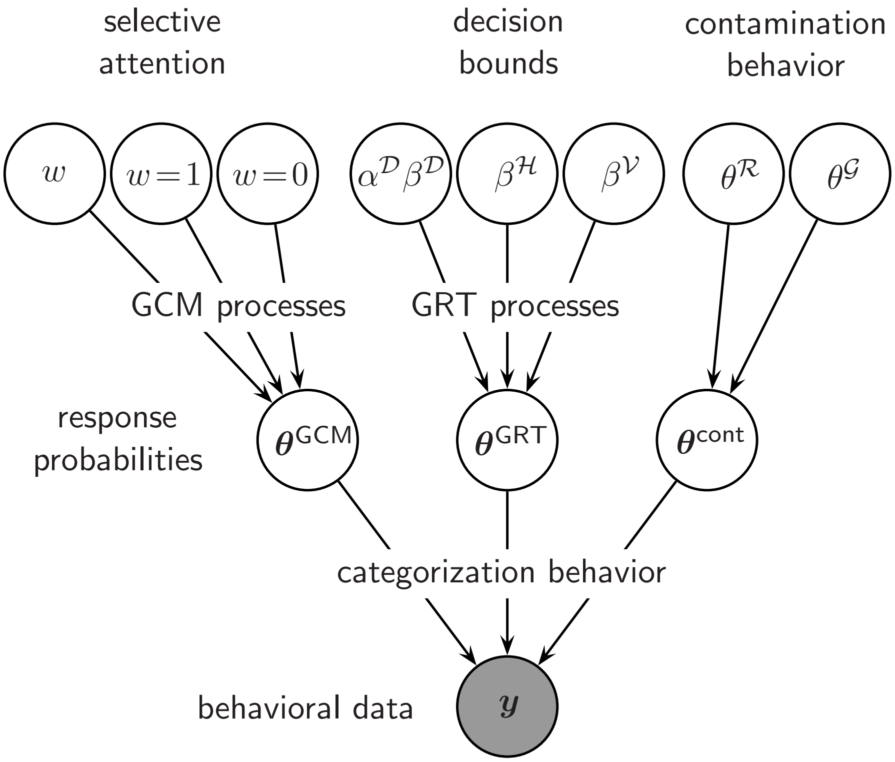


Figure 3.15: Schematic graphical model representation of the latent-mixture approach. A GCM, GRT, or contaminant categorization process generates the observed behavior of each individual. Within each model, special-case strategies involving the nature of selective attention, the decision bound, or the contaminant response probabilities are considered.

understanding how people represent categories.

3.4.1 Latent-Mixture Model

Our latent-mixture approach assumes that each subject uses one categorization model or specific strategy within that model, and that the overall data set is therefore a mixture of these specific components. We also allow for the possibility of contaminant subjects, who are guessing or repetitively assigning stimuli to the same category. Instead of filtering these people out, we model the contaminant behavior as another mixture component Zeigenfuse and Lee (2010).

Figure 3.15 presents a schematic graphical model that summarizes our approach. Each subject's categorization data $\mathbf{y}$ is generated by either the GCM, GRT, or a contaminant process. Within each of these general models, there are specific possibilities. For the GCM, either the original model with general selective attention w is used, or attention is focused on only one of the stimulus dimensions, $w = 0$ or $w = 1$. For the GRT, either a general diagonal bound is used $\alpha^{\mathcal{D}}, \beta^{\mathcal{D}}$, or unidimensional horizontal $\beta^{\mathcal{H}}$ or vertical $\beta^{\mathcal{V}}$ bounds are used. For the contaminant processes, either a category is repeatedly chosen $\theta^{\mathcal{R}}$ or a guess is made on each trial $\theta^{\mathcal{G}}$.

GCM and Strategies

The GCM is an exemplar model that assumes people store all stimuli in memory and categorize a new stimulus by comparing it the stored stimuli. It is based on a similarity comparison between the presented stimulus and every other stimulus, using the concept of psychological distance (Shepard, 1957). If the stimuli are points in a two-dimensional coordinate space, psychological distance is defined as $d_{ij} = [\sum_{k=1}^{N} w_k |x_{ik} - x_{jk}|^r]^{\frac{1}{r}}$, where x_{ik} is the value of coordinate point x_i on dimension k, N is the number of dimensions, and r is either equal to 1 or 2 for separable-dimension or integral-dimension stimuli, respectively. The selective attention parameter w controls the level of attention given to one stimulus dimension. The distance is used to calculate the similarity $\eta_{ij} = (e^{-cd_{ij}})^{\gamma}$ so as the distance between points gets larger, their perceived similarity decreases exponentially. The generalization parameter c controls the steepness of the generalization gradient. The response determinism parameter γ_k controls probabilistic or deterministic responding (Ashby and Maddox, 1993). The final probability based on these processes can be affected by the bias toward each category. In our implementation, we assume there is no bias, and so set $b = \frac{1}{2}$. This means that the probability of responding J to stimulus i is equal to $\theta^{GCM} = \sum_{j \in C_j} \eta_{ij} / \sum_{K=1}^{m} (\sum_{k \in C_K} \eta_{ik})$.

In the full GCM, selective attention can range between 0 and 1. The special cases of $w = 0$

and $w = 1$ correspond to attending to just one of the two stimulus dimensions, and constitute theoretically interesting strategies. For example, if one stimulus dimension is shape and the other is color, one person might attend only to the shape dimension and place circles in one category and squares in the other. However, another person might attend only to the color dimension and place red shapes in one category and blue shapes in the other. A third person might attend to both dimensions and categorize red circles separately from blue squares. These possibilities correspond to the three GCM components included in our model, represented by the $w = 0$, $w = 1$, and w nodes respectively.

GRT and Strategies

The GRT model assumes that instead of storing each stimulus in memory, people partition the stimulus space into response regions divided by boundaries. Response probabilities are determined by these decision bounds, based on which region a noisy perception of the presented stimulus, $x_{pi} = x_i + \epsilon_p$, belongs. Our model considers only linear decision bounds, although quadratic bounds have also been considered in the literature (Ashby and Maddox, 1992). A linear bound is defined as a discriminant function of the two dimensions satisfying the implicit line equation $h(x_1, x_2) = b_1x_1 + b_2x_2 + c$.

GRT assumes that there is variability in people's memory of the location of the bound. To account for this, the function is adjusted to include criterial noise ϵ_c. The function is compared to a bias parameter δ which captures bias toward a category. If $h(x_{pi}) + \epsilon_c$ is smaller than δ, the response is category A. If it is larger than δ, the response is category B. If it is equal to δ, the response will be a guess between A and B. We again assumed no bias so that $\delta = 0$. Thus, the probability of a category A decision for stimulus i is $\theta^{GRT} = P(h(x_{pi}) + \epsilon_c < 0)$.

Special cases of the general GRT model that have previously been emphasized involve unidimensional boundaries corresponding to vertical or horizontal lines. A vertical strategy is defined by an intercept value β^V, a horizontal strategy is defined by an intercept value, β^H, and a general diagonal boundary is defined by a slope and intercept α^D and $\beta^{\mathcal{D}}$. These possibilities correspond to the three GRT components included in our model. Although the vertical and horizontal strategies can be viewed as special cases of the diagonal strategy, one way to think about this in the latent-mixture approach is as a single model with a theoretically-rich prior. Including the vertical and horizontal boundaries as special cases corresponds to considering only a diagonal boundary with a prior that places significant density on boundaries with infinite and zero slope.

Contaminant Strategies

The three GCM strategies and the three GRT strategies make up six mixture components in our model. The remaining components capture contaminant subjects. In these cases, the probability values θ^{cont} do not follow from a theoretical model, but are set directly. For guessing, the response probability is $\theta^{\mathcal{G}} = 0.5$, so that each category response is equally likely on every trial. For a repetitive contaminant behavior, the probability of a category A response is either $\theta^{\mathcal{G}}_A = 0.99$ or $\theta^{\mathcal{G}}_B = 0.01$, depending on which category choice is repeated. Adding these three contaminant possibilities leads to a total of nine components of our latent mixture model, with Figure 3.15 combining the two repeated contaminant possibilities.

3.4.2 Modeling Results

We implemented the graphical model in JAGS (Plummer, 2003), and used fully Bayesian methods based on MCMC sampling to make inferences. Our results are based on 3 independent chains with 100 samples each, collected after discarding the first 500 burn-in ones from each chain, and testing for convergence using the standard $\hat{R}$ statistic. Advantage of

Experiment	# Subjects	# Blocks	# Stimuli	# Conditions	Type of Stimuli
Kruschke	160	8	8	4	Rectangles
Bartlema et al.	65	40	8	2	Shepard circles
Zeithamova & Maddox	170	5	80	4	Gabor patches
Navarro et al.	40	8	25	4	Faces

Table 3.2: Properties of the categorization experiments.

this methodological approach include accounting coherently for uncertainty about inferences, both in terms of model use and model-specifc parameters, and automatically controlling for the different complexity of the models and strategies considered (Lee and Wagenmakers, 2013).

We applied the model to four previously published categorization experiments. These experiments all involved a series of trials in which subjects viewed a stimulus and placed it into one of two categories, with corrective feedback after each trial. The stimuli used varied across experiments and include rectangles varying in size and interior line segment position Kruschke (1993a), Shepard circles varying in size and radial lines Bartlema et al. (2014), Gabor patches varying in frequency and orientation Zeithamova and Maddox (2006), and faces Navarro et al. (2005). For the first three, there is a natural two-dimensional stimulus representation. For the faces, we assumed a two-dimensional representation based on multidimensional scaling modeling Okada and Lee (2016). Details of the experiments, including the number of subjects, blocks, nature of the experiment, and the various conditions, are presented in Table 3.2.

Table 3.3 summarizes our results by listing how many people are inferred most likely be using each of the possible models and strategies. The individual model-use inferences come are seen in the indicator variable in the JAGS code that assumes a uniform prior over all nine potential models, meaning that in the prior, each person is equally likely to use any of the nine. The "most likely" model for each person is taken from the posterior distribution of the indicator variable. There are four conditions in the Kruschke (1993a) experiment: the

		Exemplar			Bound			Contam.	
		w	0	1	V	H	D	G	R
Kruschke	Filtration 1	10	-	30	-	-	-	-	-
	Filtration 2	6	30	1	3	-	-	-	-
	Condensation 1	15	4	5	-	4	8	4	-
	Condensation 2	19	6	10	2	-	1	2	-
	Total	**50**	**40**	**46**	**5**	**4**	**9**	**6**	**-**
Bartlema et al.	Diagonal	-	3	-	5	15	7	1	-
	Criss-Cross	-	3	4	8	8	4	7	-
	Total	**-**	**6**	**4**	**13**	**23**	**11**	**8**	**-**
Zeithamova & Maddox	Unidimensional	-	2	-	31	1	1	5	1
	Unidimensional + load	3	5	7	22	5	2	5	1
	Information-Integration	-	-	-	20	-	11	2	1
	Information-Integration + load	1	9	3	19	0	10	3	-
	Total	**4**	**16**	**10**	**92**	**6**	**24**	**15**	**3**
Navarro et al.	Gender	9	-	1	-	-	-	-	-
	Hair	3	3	1	-	3	-	-	-
	Trust	4	1	1	2	-	1	1	-
	Random	4	-	2	-	1	-	3	-
	Total	**20**	**4**	**5**	**2**	**4**	**1**	**4**	**-**

Table 3.3: Number of participants inferred to use an exemplar, decision bound, or contaminant strategy in each data set.
(w: uniform w strategy; 0: w=0 strategy; 1: w=1 strategy; V: vertical; H: horizontal; D: diagonal; G: guess; R: repeat (either all category A or all category B).

first two are filtration category structures, in which the stimuli can be categorized correctly by using information from only one dimension, and the second two are condensation category structures, in which the stimuli can only be categorized correctly by using information from both dimensions. The majority of the 160 participants are inferred to use the GCM exemplar approach, but the specific selective attention strategy varies by condition. The Bartlema et al. (2014) experiment has two conditions, named after the category structures, both of which require information from both stimulus dimensions for correct categorization. The majority of the 65 participants use a decision bound approach. The Zeithamova and Maddox (2006) experiment has four conditions. The unidimensional condition is similar to the filtration condition in the Kruschke experiment and the information-integration condition is similar to the condensation condition. The "+ load" label in Table 3.3 indicates that that condition also involved a simultaneous working memory load task. The majority of the 170 participants use the decision-bound approach, with the vertical strategy being most common. This experiment involves the most contaminant subjects, who are inferred primarily to be guessing. The Navarro et al. (2005) experiment has four conditions. These involved categorizing faces based on gender, hair color, perceived level of trust, and a random condition with no logical structure. The majority of the 40 subjects use an exemplar approach, with selective attention that considers both available dimensions, but there is large individual variation over both models and strategies across the conditions.

Kruschke (1993a) Results

The results from the Kruschke (1993a) experiment are shown at an individual level, for selected subjects, in Figure 3.16. The circles show the eight stimuli. The dark-colored circles show a response of category A while the light-colored circles show a response of category B. The size of the circle shows the number out of the total number of trials that each stimulus was placed in either category. The smallest circle means that stimulus was placed into that

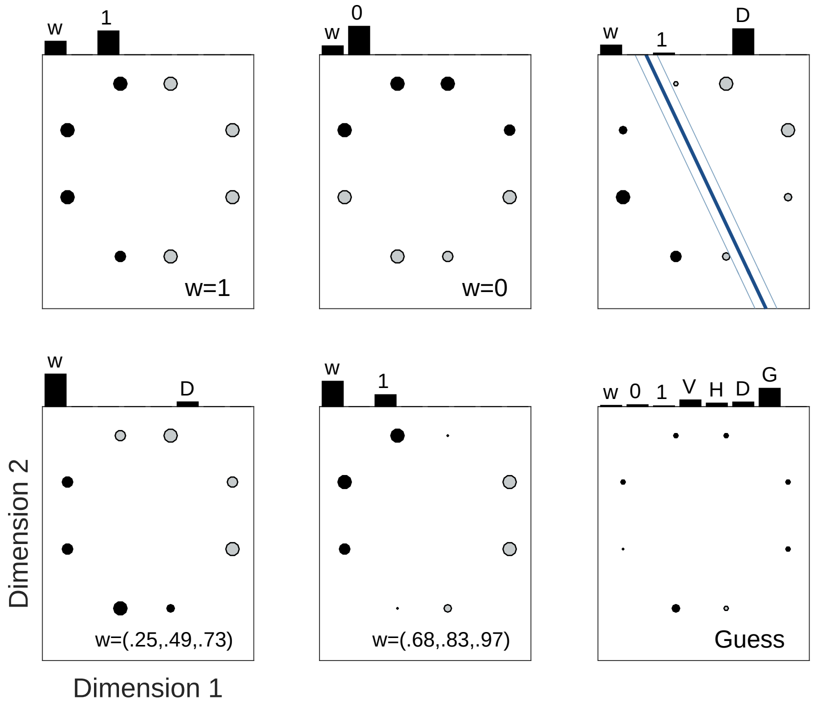


Figure 3.16: Inferred model or strategy use, and attention values or decision bounds, for selected subjects from the Kruschke (1993a) experiment.

category exactly half of the time while the largest circle means that stimulus was placed into that category all the time. The bar graphs on the top of each panel show the uncertainty in the inference about which model and strategy the subject used. Each bar shows the posterior probability for a model or strategy. A tall bar showing one strategy means that we can be more certain of that person's inferred strategy than when there are shorter bars showing many strategies. The text at the bottom right corner of each panel indicates the inferred most likely strategy. For the general GCM w strategy, the 95% credible intervals and posterior mean for w are listed. For the GRT possibilities, the bound corresponding to the posterior mean is shown as a thick line, and the upper and lower bounds f to the 95% credible intervals are shown as thin lines.

These subject in Figure 3.16 are chosen to include at least one subject from each condition. The top-left came subject from the Filtration 1 condition, the top-middle subject

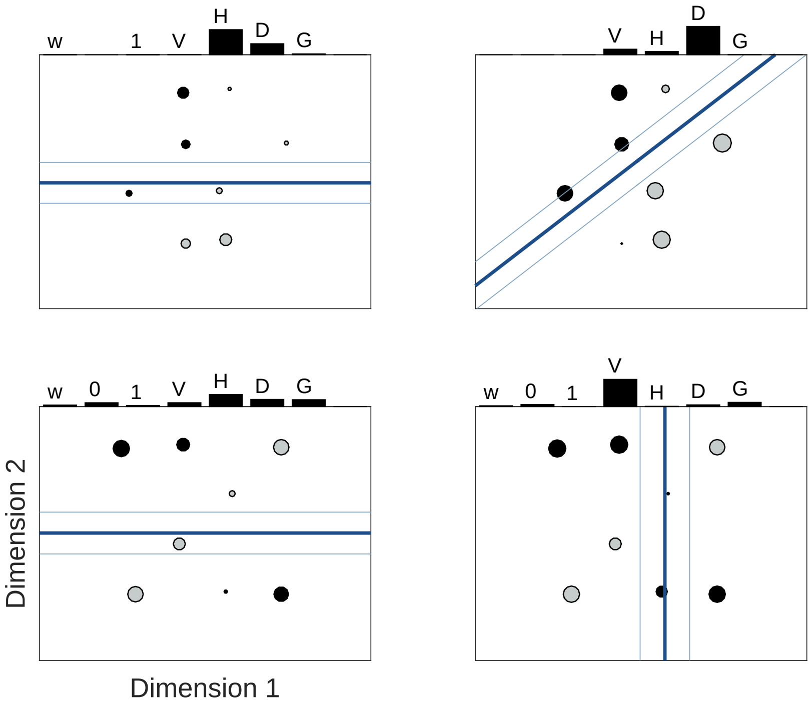


Figure 3.17: Inferred model or strategy use, and attention values or decision bounds, for selected subjects from the Bartlema et al. (2014) experiment.

came from the Filtration 2 condition, the top-right and bottom-left subjects came from the Condensation 1 condition, and the bottom-middle and bottom-right subjects came from the Condensation 2 condition. The first two subjects from the filtration conditions are inferred to be most likely using an exemplar strategy with $w = 1$ and $w = 0$, with some possibility of the general GCM w strategy. The subjects from the condensation conditions are inferred to be more likely to use either a diagonal boundary or a general GCM w strategy with a mean value close to $w = 0.5$ in one case, and $w = 0.83$ in the other. The last subject is inferred to be a guessing contaminant, with a larger degree of uncertainty.

Bartlema et al. (2014) Results

The results from the Bartlema et al. (2014) experiment are shown at an individual level in Figure 3.17. The top panels come from the diagonal condition and the bottom panels

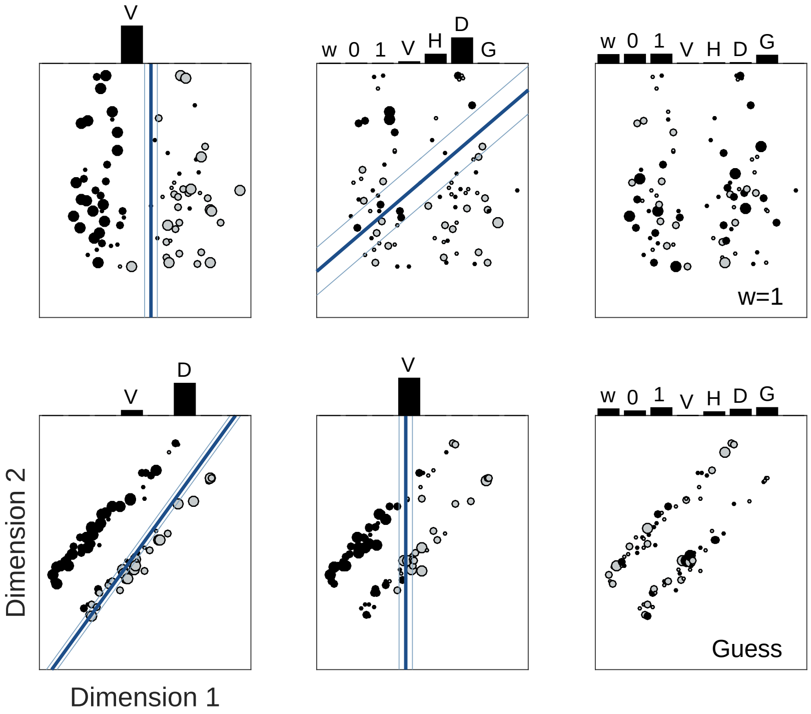


Figure 3.18: Inferred model or strategy use, and attention values or decision bounds, for selected subjects from the Zeithamova and Maddox (2006) experiment.

come from the criss-cross condition. The top-left subject is inferred to be using a horizontal boundary, but with some uncertainty about the possible use of a more general diagonal boundary. The bottom-left subject is also inferred to be using a horizontal boundary, but there is a possibility of a diagonal boundary, or a contaminant guessing strategy. The top-right subject is inferred to be using a diagonal boundary, with greater uncertainty. The bottom-right subject is inferred to be using a vertical boundary, also with a high level of certainty.

Zeithamova and Maddox (2006) Results

The results from the Zeithamova and Maddox (2006) experiment are shown at an individual level in Figure 3.18. The top-left subject comes from the unidimensional condition. The top-middle and top-right subjects come from the unidimensional + load condition. The bottom-left subject come from the information-integration condition. The bottom-middle and bottom right subjects come from the information-integration + load condition. In this experiment, very few of the subjects were inferred to be using an exemplar strategy, perhaps as a result of the large number of stimuli required to keep in memory. Even though most subjects were inferred to be using a decision bound, there is still great variation in the specific shape of the boundaries, with varying slopes and intercepts. Two of the subjects selected for Figure 3.18 are inferred to be using a vertical boundary, even though they come from conditions with different category structures. Similarly, two of the subjects are inferred to be using a diagonal boundary, but one with more uncertainty than the other about the location of the boundary. This experiment also involved a large number of subjects inferred to be contaminants, one of whom is shown in Figure 3.18. The top-right panel shows one subject who was inferred to be using an exemplar approach with a w=1 strategy, although there is large uncertainty about this inference, consistent with the poor categorization performance shown.

Navarro et al. (2005) Results

The results from the Navarro et al. (2005) experiment are shown at an individual level in Figure 3.19. The top-left subject comes from the gender condition, the top-middle and top-right subjects come from the hair-color condition, the bottom-left subject comes from the trust condition, and the bottom-middle and bottom-right subjects come from the random condition. Most of the subjects are inferred to be using the GCM, perhaps as a result of the stimuli being faces and not easily separable into psychologically interpretable dimensions. Two of the selected subjects are inferred to be using the general GCM w strategy, with varying mean values depending on the condition. A few subjects are inferred, with less certainty, to be using a decision-bound approach. The random condition has the most contaminants, as for the subject in the bottom-right panel, typically with large uncertainty

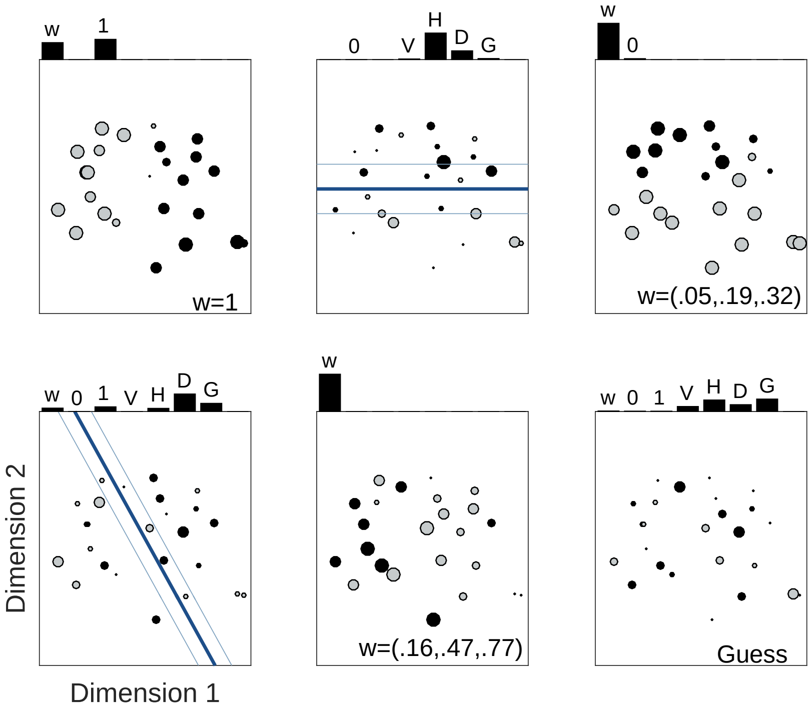


Figure 3.19: Inferred model or strategy use, and attention values or decision bounds for selected subjects from the Navarro et al. (2005) experiment.

about model use.

3.4.3 Discussion

I have presented a latent-mixture model that infers which of the two models – the GCM or the GRT – each person is using, and whether they are using a specific strategy within that model. Our individual differences analysis showed that different people's categorization behavior can best be explained by different model strategies, depending on the types and number of stimuli involved, and the nature of the category structures. Instead of continuing a debate of a "one model fits all" answer where all behavioral data must be in accordance with one type of model, applications of our modeling approach to individual subject data can potentially reveal multiple models and strategies being used by different people.

Future work could apply our general method, and the specific model we have implemented and demonstrated, to other categorization experiments, exploring how individual differences change with the type of stimuli and category structures involves. It would be interesting to understand individual differences for more complicated real-world stimuli, such as faces, with the goal of understanding how people categorize in everyday life. It is straightforward to extend our model to include other theoretical accounts of categorization behavior, and different specific strategies within them. These could incorporate other categorization models, such as hybrid models that combine prototype with exemplar or rule-based representations. It would also be possible to extend the model to allow for shifts in categorization within an individual, allowing for possibilities like rapid shifts in attention, or the adaptation of an overly simple unidimensional bound to a more general diagonal bound on the basis of feedback. Examination of the strategy shifts that occur can be useful for further predictive modeling of when we can expect participants to switch strategies. Collectively, these extensions allow for broader and deeper investigation of the individual differences in the way people represent and use categories.

3.5 Conclusion

Moving forward, one attraction of our modeling approach is its generality. It is possible for both the current case studies, and for other case studies — involving other stimuli or category structures — that quite different categorization models will be appropriate. For example, some category structures will need nonlinear decision bounds in the GRT, and individual differences in generalization gradients in the GCM will be needed for stimuli that allow for perceptual learning. Beyond the GRT and GCM there are many other theories and models of categorization, including ALCOVE, COVIS, RULEX, SUSTAIN, and hybrid models (Kruschke, 1992; Ashby et al., 1998; Nosofsky and Palmeri, 1998; Love et al., 2004;

Busemeyer et al., 1984; Smith and Minda, 2000), that could be used as the underlying psychological models in our wisdom of the crowd framework.

As well as considering other models of categorization, our approach would benefit from extended models of categorization decisions and category learning. For example, it is possible that people change strategies during the course of learning a category structure. In the GCM case study, some participants change how they attend to the different stimulus dimension over the course of learning. These changes are not formally part of the GCM model that we used, nor is it a common capability in most established psychological models of category learning. It would also be possible to extend the modeling approach to allow for individual differences in terms of which psychological model each person uses. It may be that some people use decision bounds while others use exemplar-based similarity, and it may even be that some people start with an exemplar strategy and switch to a decision bound strategy as the number of presented stimuli increases. Both of these extensions could be naturally accommodated by hierarchical and latent-mixture extensions within the graphical modeling approach we have used, and could continue to be applied to data using Bayesian methods.

Turning to applied possibilities, a challenge for our approach is determining how to represent the stimuli. The simple perceptual nature of the Gabor patch stimuli is not true of all stimuli, and the representation of the faces that we used was based on independent similarity data collection and multidimensional scaling analysis. Even then, the representation of the face stimuli only applied to 25 faces, and we have no method for determining how a new face should be represented in this same space. What is needed for real-world application is a formal method for determining an appropriate representation of any possible stimulus. To return to our motivating example of doctors learning to diagnose skin patches, it seems possible, but far from trivial, that image processing methods could automatically map a visual skin patch stimulus into a dimensional psychological representation. In general, the applicability of our approach to real-world situations hinges on finding such a representational method. When such methods are available, our approach has the attractive property of requiring relatively limited effort on the part of people to categorize large numbers of stimuli. Once a categorization model has been inferred for each individual, it can be applied to any number of new stimuli. The accuracy of the crowd categorizations should increase as both more individuals are included in the group, and as individuals categorize more stimuli.

One way to interpret our wisdom of the crowd approach comes from machine learning where it would be called boosting (Hastie et al., 2001). Under this interpretation, the model of each individual functions as a weak classifier and there is a simple majority rule for aggregation of the categorization decision. In fact, the GRT is closely related to decision-bound methods like support vector machines, and the GCM is closely related to radial basis classifiers, nearest-neighbor, and other clustering methods (Welinder et al., 2010; Gomes et al., 2011). From a machine learning perspective, the contribution of our approach is to help identify useful weak classifiers, by recognizing that the classification problem is a problem of human categorization, and so domain-specific cognitive models should be effective in ways that more domain-general statistical methods may not. Nevertheless, it is almost certainly possible to improve categorization accuracy in the case studies we have presented using established and successful machine learning techniques. In particular, it is likely that discriminative machine learning methods could outperform the generative approach to probabilistic modeling we have used. The strength of the psychological nature of our approach comes not from relative accuracy, but from significantly greater interpretability. A recognized challenge for machine learning methods relates to issues of interpretability and trust (Ribeiro et al., 2016). Whereas a deep neural net may only be able to give a sequence of connection weights as a justification for a decision, it is generally easy to give complete and meaningful accounts of how and why our aggregated crowds decided to categorize a new stimulus a certain way. These explanations will reference interpretable decision strategies and individual differences in those strategies. This should not only increase the probability that people trust the crowd decision, but also make training and remediation of individuals possible, especially by comparing their categorization strategies to others.

In terms of psychological understanding, our approach is a good example of what Watts (2017) calls "solution-oriented" social science. The general goal is to seek to solve a practical problem, using existing theories and models where possible, and identifying gaps where they exist. In our case, the wisdom of the crowd problem demands that the modeling of individual differences be taken seriously, and both of our case studies incorporated different sorts of categorization strategies as well as allowing for different types of contaminant behavior. This is relatively new theoretical territory for the modeling of human category learning, and there certainly is not wide exploration or agreement on the number and type of these individual differences. In this way, our results provide new empirical evaluation and are relevant to the development of theory. More tellingly, the results for the faces case study identify the need for models of how people change or adapt their categorization strategies over the course of learning. There are few such theories, and no established categorization models that include this capability. In these sorts of ways, our case studies not only demonstrate the applicability of current categorization models to have a useful real-world application, but highlight the role of applications in focusing attention of important theoretical and modeling problems that need to be solved to understand how categorization works.

Chapter 4

Conclusion

The wisdom of the crowd phenomenon has been demonstrated in a variety of domains, including simple numerical estimates, probability judgments of real-world events, predictions of sports events, and now categorization tasks. These tasks fulfill all four requirements of a wise crowd (Surowiecki, 2004). They show *diversity* in allowing the decision-makers to have a wide range of different opinions and backgrounds, such as varying rates of learning or different degrees of mis-calibration of probabilities. The tasks are *decentralized* in that the decision-makers all draw their estimates from different sources, such as different expertise levels. The tasks show *independence* by requiring that the decision-makers produce their answers by themselves. Finally, they all rely on effective methods of *aggregation*, which I have demonstrated as either simple statistical measures of the mean, median, and mode or as more sophisticated cognitive models that tap into people's latent decision-making processes.

In chapter 2, I introduced a hierarchical cognitive model for combining people's estimates of probabilities, in an environment where the ground truths were already known, as well as in a truly predictive environment in which the model predicted winning percentages of sports teams for seasons that had not yet happened at the time of the study. Apart from performing fairly well (relative to other aggregation measures) in accurately predicting probabilities, I demonstrated how we also learned about the participants' decision-making processes and individual differences. I showed how we identified experts in the survey, even when the model was naive to the truth. I also showed how we learned about the participants' misperception of probabilities, in their under- and over-estimations of extreme probabilities.

In chapter 3, I explored the applications of the wisdom of the crowd for the field of category learning. I showed how we used a number of existing categorization data sets to prove that this phenomenon exists in this field if you combine decisions via the modal responses. I then introduced a Bayesian framework for the GRT model of categorization so that we could compare its inferences with the GCM in a more straightforward manner. I showed how a combination of the cognitive models and the modal response measure correctly categorized newly generated stimuli that have not been seen by the participants. As in the previous chapter, I demonstrated how we learned about individual differences in our participants, in the use of boundaries, selective attention, and strategy use.

Overall, the work presented here provides a deeper understanding of the benefit of cognitive modeling for aggregating people's decisions. This approach not only often provides improved prediction performance for tasks like probability estimation and categorization, but it also provides a psychological insight into human behavior. Being able to account for individual differences in strategies, expertise, or biases will lead to even better cognitive models, which in turn will not only help us understand more about how people generate decisions but will also lead to more accurate crowd answers in predictive applications.

Bibliography

J Albert, J Bennett, and J J Cochran, editors. *Anthology of statistics in sports*. SIAM, 2005.

F. G. Ashby and R. E. Gott. Decision rules in the perception and categorization of multidimensionalstimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:33–53, 1988.

F. G. Ashby and W. T. Maddox. Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):50–71, 1992.

F. G. Ashby and W. T. Maddox. Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37(3):372–400, 1993.

F. G. Ashby and J. T. Townsend. Varieties of perceptual independence. *Psychological Review*, 93:154–79, 1986.

F. G. Ashby, L. A. Alfonso-Reese, A. U. Turken, and E. M. Waldon. A neuropsycholoigcal theory of multiple systems in category learning. *Psychological Review*, 105:442–481, 1998.

M Bar-Hillel, D V Budescu, and M Amar. Predicting world cup results: Do goals seem more likely when they pay off? *Psychonomic Bulletin & Review*, 15:278–283, 2008.

A. Bartlema. [Selective attention in category learning] Unpublished raw data. 2013.

A. Bartlema, M.D. Lee, R. Wetzels, and W. Vanpaemel. A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, 59:132–150, 2014.

L A Brenner, D J Koehler, V Liberman, and A Tversky. Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavioral and Human Decision Processes*, 65:212–219, 1996.

D V Budescu and T R Johnson. A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making*, 6:857–869, 2011.

Jerome R Busemeyer, Gerald I Dewey, and Douglas L Medin. Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:638, 1984.

D R Cavagnaro, M A Pitt, R Gonzalez, and I J Myung. Discriminating among probability weighting functions with adaptive design optimization. *Journal of Risk and Uncertainty.*, in press.

I Danileiko, M D Lee, and M. Kalish. A Bayesian latent mixture approach to modeling individual differences in categorization using General Recognition Theory. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, 2015.

S W Ell and F G Ashby. The impact of category separation on unsupervised categorization. *Attention, Perception & Psychophysics*, 74:466–475., 2012.

A. Gelman. Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 131–143. Chapman & Hall/CRC, Boca Raton (FL), 1996.

A Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006.

W M Goldstein and H J Einhorn. Expression theory and the preference reversal phenomena. *Psychological Review*, 94:236–254, 1987.

Ryan G Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. Crowdclustering. In *Advances in Neural Information Processing Systems*, pages 558–566, 2011.

R Gonzalez and G Wu. On the shape of the probability weighting function. *Cognitive Psychology*, 38:129–166., 1999.

R Hastie and T Kameda. The robust beauty of majority rules in group decisions. *Psychological Review*, 112:494–508, 2005.

T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning.* Springer Verlag, New York, NY, 2001.

S M Herzog and R Hertwig. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20:231–237, 2009.

M I Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.

D. Koller, N. Friedman, L. Getoor, and B. Taskar. Graphical models in a nutshell. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning.* MIT Press, Cambridge, MA, 2007.

J K Kruschke. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44, 1992.

J K Kruschke. Human category learning: Implications for backpropagation models. *Connection Science*, 5:3–36, 1993a.

J K Kruschke. Three principles for models of category learning. *The Psychology of Learning and Motivation*, 29:57–90, 1993b.

M D Lee. How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55:1–7, 2011.

M D Lee. Bayesian methods in cognitive modeling. In *The Stevens Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition.* John Wiley & Sons, in press.

M D Lee and D J Navarro. Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9:43–58, 2002.

M D Lee and E.-J. Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2013.

M D Lee, S Zhang, and J Shi. The wisdom of the crowd playing the Price is Right. *Memory & Cognition*, 39:914–923, 2011.

M D Lee, M Steyvers, M de Young, and B J Miller. Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*, 4:151–163, 2012.

M D Lee, M Steyvers, and B J Miller. A cognitive model for aggregating people's rankings. *PLoS ONE*, 9:1–9, 2014.

Michael D Lee and Irina Danileiko. Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9(3):259, 2014.

S Lewandowsky. Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37: 720, 2011.

S Lichtenstein, B Fischoff, and L D Phillips. Calibration of probabilities: The state of the art to 1980. In D Kahneman, P Slovic, and A Tversky, editors, *Judgment under uncertainty: Heuristics and biases*, pages 306–334. Cambridge University Press, Cambridge, England, 1982.

Bradley C Love, Douglas L Medin, and Todd M Gureckis. SUSTAIN: a network model of category learning. *Psychological Review*, 111:309, 2004.

R D Luce. Detection and recognition. In R D Luce, R R Bush, and E Galanter, editors, *Handbook of Mathematical Psychology*, pages 103–189. Wiley, New York, NY, 1963.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modelling framework: Concepts, structure and extensibility. *Statistics and Computing*, 10:325–337, 2000.

W. T. Maddox and F. G. Ashby. Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, 53:49–70, 1993.

E C Merkle. Calibrating subjective probabilities using hierarchical Bayesian models. In *Social Computing, Behavioral Modeling, and Prediction (SBP) 2010*, volume 6007 of *Lecture Notes in Computer Science*, pages 13–22. 2010.

E C Merkle and M Steyvers. A psychological model for aggregating judgments of magnitude. *Lecture Notes in Computer Science*, 6589:236–243, 2011.

D J Navarro, M D Lee, and H Nikkerud. Learned categorical perception for natural faces. In B G Bara, L W Barsalou, and M Bucciarelli, editors, *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1600–1605. Erlbaum, Mahwah, NJ, 2005.

R M Nosofsky. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):104–114, 1984.

R M Nosofsky. Attention, similarity and the idenitification-categorization relationship. *Journal of Experimental psychology: General*, 115:39–57, 1986.

R M Nosofsky. Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43:25–53, 1992.

R M Nosofsky and T J Palmeri. A rule-plus-exception model for classifying objects in continuous-dimensionspaces. *Psychonomic Bulletin & Review*, 5:345–369, 1998.

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

M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Vienna, Austria, 2003.

D Prelec. The probability weighting function. *Econometrica*, 66:497–527, 1998.

S. K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3(3):382–407, 1972.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016.

R Selker, M D Lee, and R Iyer. Thurstonian cognitive models for aggregating top-n lists. *Decision*, in press.

R N Shepard. Stimulus and response generalization: A stochastic model relating generalizationto distance in psychological space. *Psychometrika*, 22(4):325–345, 1957.

R M Shiffrin, M D Lee, W.-J. Kim, and E.-J. Wagenmakers. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32:1248–1284, 2008.

J. D. Smith and J. P. Minda. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24:1411–1436, 1998.

J D Smith and J P Minda. Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:3–27, 2000.

J. D. Smith and J. P. Minda. Journey to the center of the category: the dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(4):984–1002, 2001.

J. D. Smith and J. P. Minda. Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28:800–11, 2002.

R D Sorkin, C J Hays, and R West. Signal-detection analysis of group decision-making. *Psychological Review*, 108:183–203, 2001.

Fabian A Soto, Lauren Vucovich, Robert Musgrave, and F Gregory Ashby. General recognition theory with individual differences: a new method for examining perceptual and decisional interactions with an application to face perception. *Psychonomic Bulletin & Review*, pages 1–24, 2015.

J Surowiecki. *The Wisdom of Crowds*. Random House, New York, 2004.

Brandon M Turner, Mark Steyvers, Edgar C Merkle, David V Budescu, and Thomas S Wallsten. Forecast aggregation via recalibration. *Machine learning*, 95(3):261–289, 2014.

A Tversky and D Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.

Amos Tversky and Craig R Fox. Weighing risk and uncertainty. *Psychological review*, 102 (2):269, 1995.

L Ungar, B Mellors, Satopää, J Baron, P Tetlock, J Ramos, and S Swift. The good judgment project: A large scale test of different methods of combining expert predictions. Technical report, AIII Technical Report FS-12-06. Machine Aggregation of Human Judgment, 2012.

W Vanpaemel. BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods*, 41:1111–1120, 2009.

W Vanpaemel and M D Lee. Using priors to formalize theory: Optimal attention and the Generalized Context Model. *Psychonomic Bulletin & Review*, 19:1047–1056, 2012.

E Vul and H Pashler. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19:645–647, 2008.

Duncan J Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1:1–5, 2017.

D J Weiss and J Shanteau. Who's the best? A relativistic view of expertise. *Applied Cognitive Psychology*, in press.

Peter Welinder, Steve Branson, Serge J Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *NIPS*, volume 23, pages 2424–2432, 2010.

J F Yates. *Judgment and decision making.* Prentice Hall, Englewood Cliffs, NJ, 1990.

Sheng Kung Michael Yi, Mark Steyvers, Michael D Lee, and Matthew J Dry. The wisdom of the crowd in combinatorial problems. *Cognitive Science*, 36:452–470, 2012.

M D Zeigenfuse and M D Lee. Finding the features that represent stimuli. *Acta Psychologica*, 133:283–295, 2010.

D Zeithamova and W T Maddox. Dual task interference in perceptual category learning. *Memory & Cognition*, 34:387–398, 2006.

Hang Zhang and Laurence T. Maloney. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action and cognition. *Frontiers in Neuroscience*, 6:1–14, 2012.